Springer Handbook of Mechanical Engineering

Grote
Antonsson
Editors

Springer

Springer Handbook of Mechanical Engineering

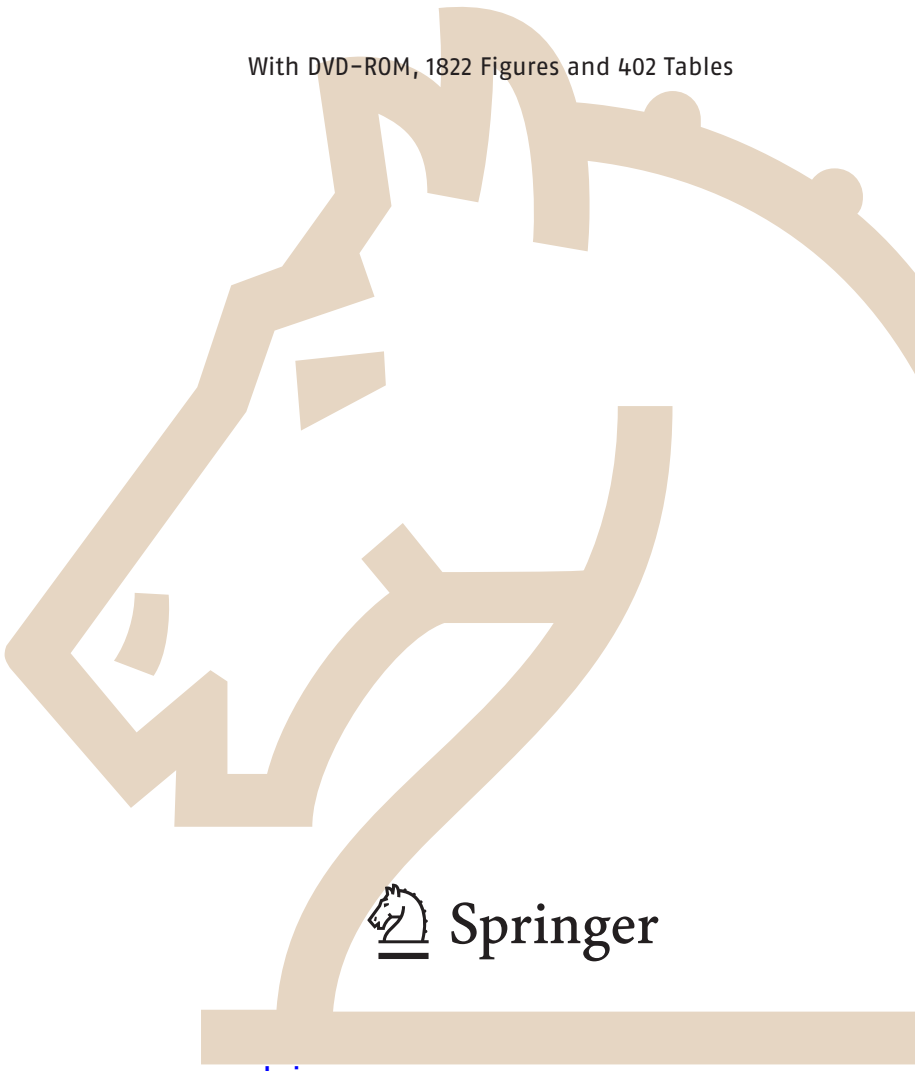
Springer Handbooks provide a concise compilation of approved key information on methods of research, general principles, and functional relationships in physical sciences and engineering. The world's leading experts in the fields of physics and engineering will be assigned by one or several renowned editors to write the chapters comprising each volume. The content is selected by these experts from Springer sources (books, journals, online content) and other systematic and approved recent publications of physical and technical information.

The volumes are designed to be useful as readable desk reference books to give a fast and comprehensive overview and easy retrieval of essential reliable key information, including tables, graphs, and bibliographies. References to extensive sources are provided.

Springer Handbook of Mechanical Engineering

Grote, Antonsson (Eds.)

With DVD-ROM, 1822 Figures and 402 Tables



Springer

Editors:

Professor Dr.-Ing. Karl-Heinrich Grote
Department of Mechanical Engineering
Otto-von-Guericke University Magdeburg
Universitätsplatz 2
39106 Magdeburg, Germany
karl.grote@ovgu.de

Professor Erik K. Antonsson
Department of Mechanical Engineering
California Institute of Technology (CALTEC)
1200 East California Boulevard
Pasadena, CA 91125, USA
erik@design.caltech.edu

Library of Congress Control Number: 2008934575

ISBN: 978-3-540-49131-6 e-ISBN: 978-3-540-30738-9

Production and typesetting: le-tex publishing services oHG, Leipzig
Senior Manager Springer Handbook: Dr. W. Skolaut, Heidelberg
Illustrations: schreiberVIS, Seeheim and Hippmann GbR, Schwarzenbruck
Cover design: eStudio Calamar Steinen, Barcelona
Cover production: WMXDesign GmbH, Heidelberg
Printing and binding: Stürtz GmbH, Würzburg

Printed on acid free paper

SPIN 10934364 60/3180/YL 5 4 3 2 1 0

Preface

Mechanical engineering is a broad and complex field within the world of engineering and has close relations to many other fields. It is an important economic factor for all industrialized countries and the global market allows for wide international competition for products and processes in this field. To stay up to date with scientific findings and to apply existing knowledge in mechanical engineering it is important to renew and continuously update existing information.

The editors of this *Springer Handbook on Mechanical Engineering* have worked successfully with 92 authors worldwide to include chapters about all relevant mechanical engineering topics. However, this Handbook cannot claim to cover every aspect or detail of the mechanical engineering areas or fields included, and where mechanical engineers are currently present and contributing their expertise and knowledge towards the challenges of a better world. However, this Handbook will be a valuable guide for all who design, develop, manufacture, operate, and use mechanical artefacts.

We also hope to spark interest in the field of mechanical engineering from others. In this Handbook, high-school students can get a first glance at the options in this field and possible career moves.

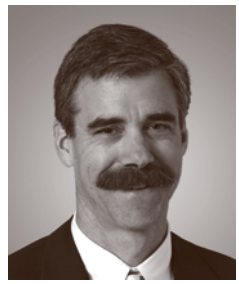
We, the editors, would like to express our gratitude and thanks to all of the authors of this Handbook, who have devoted a considerable amount of time towards this project. We would like to thank them for their patience and cooperation, and we hope for a long-lasting partnership in this ambitious project. We would also most sincerely like to thank our managers and friends at Springer and le-tex. The executives at Springer–Verlag were always most cooperative and supportive of this Handbook. Without Dr. Skolaut's continuous help and encouragement and Ms. Moebes' and Mr. Wieczorek's almost daily requests for corrections, improvements, and progress reports it would have taken another few years – if ever – to publish this Handbook. Stürtz has done a fantastic job in printing and binding. Finally we would like to thank all the people we work with in our departments and universities, who tolerated the time and effort spent on this book.

Finally, we know that there is always room for improvement – with this Handbook as with most engineering products and approaches. We, as well as the authors welcome your fair hints, comments, and criticism. Through this Handbook and with the authors' efforts, we would also like to draw your attention to what has been accomplished for the benefit of the engineering world and society.

Berlin, Fall 2008 Karl-Heinrich Grote
Pasadena, Fall 2008 Erik K. Antonsson

About the Editors

Dr. Karl-Heinrich Grote is a Professor and Chair of the Department of Mechanical Engineering – Engineering Design at the Otto-von-Guericke University in Magdeburg, Germany. He earned his “Diploma in Mechanical Engineering” (Masters of Science in Mechanical Engineering) in 1979 and his “Dr.-Ing.” (Ph.D. in Engineering) in 1984, both from the Technical University in Berlin, Germany. After a post doctoral stay in the USA he joined an automotive supplier as manager of the engineering design department. In 1990 he followed a call to become full professor at the Mechanical Engineering Department at the California State University, Long Beach, USA. In 1992 he received the TRW Outstanding Faculty award and in 1993 the VDI "Ring of Honor" for his research on Engineering Design and Methodology. In 1995 he was named chair of the Engineering Design Department at the Otto-von-Guericke University in Magdeburg, where he is now Dean of the College of Mechanical Engineering. From October 2002 to September 2004 he was Visiting Professor of Mechanical Engineering at the California Institute of Technology (Caltech) USA. Since 1995 he is Editor of the DUBBEL (Taschenbuch für den Maschinenbau) and author of several books.



Dr. Erik Antonsson is a Professor of Mechanical Engineering at the California Institute of Technology in Pasadena, where he organized the Engineering Design Research Laboratory and has conducted research and taught since 1984. He earned a Bachelor of Science in Mechanical Engineering from Cornell University in 1976, and a PhD in Mechanical Engineering from the Massachusetts Institute of Technology, Cambridge in 1982. In 1984 he joined the Mechanical Engineering Faculty at the California Institute of Technology, where he served as the Executive Officer (Chair) from 1998 to 2002. From September, 2002 through January, 2006, Dr. Antonsson was on leave from Caltech and served as the Chief Technologist at NASA’s Jet Propulsion Laboratory (JPL). He was an NSF Presidential Young Investigator (1986-1992), won the 1995 Richard P. Feynman Prize for Excellence in Teaching, and was a co-winner of the 2001 TRW Distinguished Patent Award. Dr. Antonsson is a Fellow of the ASME, and a member of the IEEE, AIAA, SME, ACM, and ASEE. He has published over 110 scholarly papers in the field of engineering design research, has edited two books, and holds eight U.S. patents.

List of Authors

Gritt Ahrens
Daimler AG X944
Systems Integration and Comfort Electric
71059 Sindelfingen, Germany
e-mail: *gritt.ahrens@daimler.com*

Seddik Bacha
Université Joseph Fourier
Grénoble Electrical Engineering Laboratory
Saint Martin d'Hères
38402 Grenoble, France
e-mail: *seddik.bacha@g2elab.inpg.fr*

Stanley Baksi
TRW Automotive, Lucas Varity GmbH
Carl Spaeter Str. 8
56070 Koblenz, Germany
e-mail: *stanley.baksi@trw.com*

Thomas Böllinghaus
Federal Institute for Materials Research
and Testing (BAM)
Unter den Eichen 87
12205 Berlin, Germany
e-mail: *thomas.boellinghaus@bam.de*

Alois Breiing
Eidgenössische Technische Hochschule Zürich (ETH)
Institut für mechanische Systeme (IMES)
Zentrum für Produkt-Entwicklung (ZPE)
ETH Zentrum, CLA E 17.1, Tannenstrasse 3
8092 Zurich, Switzerland
e-mail: *breiing@imes.mavt.ethz.ch*

Eugeniusz Budny
Institute of Mechanized Construction and Rock
Mining
Racjonalizacji 6/8
02-673 Warsaw, Poland
e-mail: *e.budny@imbigs.org.pl*

Gerry Byrne
University College Dublin
School of Electrical, Electronic
and Mechanical Engineering
Belfield, Dublin 4, Ireland
e-mail: *gerald.byrne@ucd.ie*

Boris Ilich Cherpakov (deceased)

Edward Chlebus
Wrocław University of Technology
Centre for Advanced Manufacturing Technologies
Lukasiewicza 5
50-371 Wrocław, Poland
e-mail: *edward.chlebus@pwr.wroc.pl*

Mirosław Chłosta
IMBiGS – Institute for Mechanized Construction
and Rock Mining (IMBiGS)
ul. Racjonalizacji 6/8
02-673 Warsaw, Poland
e-mail: *m.chlosta@imbigs.org.pl*

Norge I. Coello Machado
Universidad Central "Marta Abreu" de Las Villas
Faculty of Mechanical Engineering
Santa Clara, 54830, Cuba
e-mail: *norgec@uclv.edu.cu*

Francesco Costanzo
Alenia Aeronautica
Procurement/Sourcing Management Department
Viale dell'Aeronautica
Pomigliano (NA), Italy
e-mail: *fcostanzo@inwind.it*

Carl E. Cross
Federal Institute for Materials Research
and Testing (BAM)
Joining Technology
Unter den Eichen 87
12200 Berlin, Germany
e-mail: *carl-edward.cross@bam.de*

Frank Dammel
Technical University
Department of Mechanical Engineering/Institute of Technical Thermodynamics
Petersenstr. 30
64287 Darmstadt, Germany
e-mail: *dammel@ttd.tu-darmstadt.de*

Jaime De La Ree
Virginia Tech
Electrical and Computer Engineering Department
340 Whittemore Hall
Blacksburg, VA 24061, USA
e-mail: *jreelope@vt.edu*

Torsten Dellmann
RWTH Aachen University
Department of Rail Vehicles and Materials-Handling Technology
Seffenter Weg 8
52074 Aachen, Germany
e-mail: *torsten.dellmann@ifs.rwth-aachen.de*

Berend Denkena
Leibniz University Hannover
IFW – Institute of Production Engineering and Machine Tools
An der Universität 2
30823 Garbsen, Germany
e-mail: *denkena@ifw.uni-hannover.de*

Ludger Deters
Otto-von-Guericke University
Institute of Machine Design
Universitätsplatz 2
39016 Magdeburg, Germany
e-mail: *ludger.deters@ovgu.de*

Ulrich Dilthey
RWTH Aachen University
ISF Welding and Joining Institute
Pontstr. 49
52062 Aachen, Germany
e-mail: *di@isf.rwth-aachen.de*

Frank Engelmann
University of Applied Sciences Jena
Department of Industrial Engineering
Carl-Zeiss-Promenade 2
07745 Jena, Germany
e-mail: *frank.engelmann@fh-jena.de*

Ramin S. Esfandiari
California State University
Department of Mechanical & Aerospace Engineering
Long Beach, CA 90840, USA
e-mail: *esfandi@csulb.edu*

Jens Freudenberger
Leibniz-Institute for Solid State and Materials Research Dresden
Department for Metal Physics
P.O. Box 270116
01171 Dresden, Germany
e-mail: *j.freudenberger@ifw-dresden.de*

Stefan Gies
RWTH Aachen University
Institute for Automotive Engineering
Steinbachstr. 7
52074 Aachen, Germany
e-mail: *gies@ika.rwth-aachen.de*

Joachim Göllner
Otto-von-Guericke University
Institute of Materials and Joining Technology
Department of Mechanical Engineering
Universitätsplatz 2
39016 Magdeburg, Germany
e-mail:
joachim.goellner@mb.uni-magdeburg.de

Timothy Gutowski
Massachusetts Institute of Technology
Department of Mechanical Engineering
Cambridge, MA 02139, USA
e-mail: *gutowski@mit.edu*

Takeshi Hatsuzawa
Tokyo Institute of Technology
Precision and Intelligence Laboratory
4259-R2-6, Nagatsuta-cho
226-8503 Yokohama, Japan
e-mail: *hat@pi.titech.ac.jp*

Markus Hecht
Berlin University of Technology
Institute of Land and Sea Transport Systems
Department of Rail Vehicles
Salzufer 17–19
10587 Berlin, Germany
e-mail: *markus.hecht@tu-Berlin.de*

Hamid Hefazi
California State University
Mechanical and Aerospace Engineering
Department of Mechanical and Aerospace Engineering
1250 Bellflower Boulevard
Long Beach, CA 90840, USA
e-mail: *hefazi@csulb.edu*

Martin Heilmaier
Technical University
Department of Physical Metallurgy
Petersenstr. 23
64287 Darmstadt, Germany
e-mail: *m.heilmaier@phm.tu-darmstadt.de*

Rolf Henke
RWTH Aachen University
Institute of Aeronautics and Astronautics
Wuellnerstr. 7
52062 Aachen, Germany
e-mail: *henke@ilr.rwth-aachen.de*

Klaus Herfurth
Industrial Advisor
Am Wiesengrund 34
40764 Langenfeld, Germany
e-mail: *klaus.herfurth@t-online.de*

Horst Herold (deceased)

Chris Oliver Heyde
Otto-von-Guericke University
Electric Power Networks and Renewable Energy Sources
Universitätsplatz 2
39106 Magdeburg, Germany
e-mail: *chris.heyde@ovgu.de*

Andrew Kaldos
AKM Engineering Consultants
31 Tudorville Road
Bebington, Wirral CH632 HT, UK
e-mail: *andrew.kaldos@ntlworld.com*

Yuichi Kanda
Toyo University
Department of Mechanical Engineering
Advanced Manufacturing Engineering Laboratory
2100 Kujirai
350-8585 Kawagoe-City, Japan
e-mail: *cimkanda@toyonet.toyo.ac.jp*

Thomas Kannengiesser
Federal Institute for Materials Research and Testing (BAM)
Joining Technology
Unter den Eichen 87
12200 Berlin, Germany
e-mail: *thomas.kannengiesser@bam.de*

Michail Karpenko
New Zealand Welding Centre
Heavy Engineering Research Association (HERA)
17–19 Gladding Place
Manukau City, New Zealand
e-mail: *michail.karpenko@hera.org.nz*

Bernhard Karpuschewski
Otto-von-Guericke University
Department of Manufacturing Engineering
Universitätsplatz 2
39106 Magdeburg, Germany
e-mail: *karpu@mb.uni-magdeburg.de*

Toshiaki Kimura
Japan Society for the Promotion of Machine Industry (JSPMI)
Production Engineering Department
Technical Research Institute
1-1-12, Hachiman-cho
203-0042 Tokyo, Japan
e-mail: *kimura@tri.jspmi.or.jp*

Dwarkadas Kothari
VIT University
School of Electrical Sciences
Vellore, TN 632 014, India
e-mail: *dkothari@ces.iitd.ac.in*

Hermann Kühnle
Otto-von-Guericke University
Institute of Ergonomics
Factory Operations and Automation
Universitätsplatz 2
39106 Magdeburg, Germany
e-mail: *hermann.kuehnle@ovgu.de*

Oleg P. Lelikov
Bauman Moscow State Technical University
2-nd Baumanskaya, 5
Moscow, 105005, Russia

Andreas Lindemann
Otto-von-Guericke University
Institute for Power Electronics
Universitätsplatz 2
39106 Magdeburg, Germany
e-mail: *andreas.lindemann@ovgu.de*

Bruno Lisanti
AST
Via Dante Alighieri 57
Lonate Pozzolo (VA), Italy
e-mail: *bruno.lisanti@ast-italia.com*

Manuel Marya
Schlumberger Reservoir Completions
Material Engineering
14910 Airline Road
Rosharon, TX 77583, USA
e-mail: *mmarya@slb.com*

Surendar K. Marya
GeM-UMR CNRS 6183, Ecole Centrale Nantes
Institut de Recherche en Génie Civil et Mécanique
1 Rue de la Noë
44321 Nantes, France
e-mail: *surendar.marya@ec-nantes.fr*

Ajay Mathur
Simon India Limited
Plant Engineering
Devika Tower, 6 Nehru Place
New Delhi, India
e-mail: *avm2k@vsnl.net*

Klaus-Jürgen Matthes
Chemnitz University of Technology
Institute for Manufacturing/Welding Technology
Reichenhainer Str. 70
09126 Chemnitz, Germany
e-mail: *schweisstech@mb.tu-chemnitz.de*

Henning Jürgen Meyer
Technische Universität Berlin
Berlin Institute of Technology
Konstruktion von Maschinensystemen
Straße des 17. Juni 144
10623 Berlin, Germany
e-mail: *henning.meyer@tu-berlin.de*

Klaus Middeldorf
DVS – German Welding Society
Düsseldorf, Germany
e-mail: *klaus.middeldorf@dvs-hg.de*

Gerhard Mook
Otto-von-Guericke University
Department of Mechanical Engineering
Institute of Materials and Joining Technology
and Materials Testing
Universitätsplatz 2
39016 Magdeburg, Germany
e-mail: *mook@mb.uni-magdeburg.de*

Jay M. Ochterbeck
Clemson University
Department of Mechanical Engineering
Clemson, SC 29634-0921, USA
e-mail: *jochter@clemson.edu*

Joao Fernando G. Oliveira
University of São Paulo
Department of Production Engineering
Av. Trabalhador Sãocarlense, 400
São Carlos, SP 13566-590, Brazil
e-mail: *jfgo@sc.usp.br, presidencia@ipt.br*

Antje G. Orths
Energinet.dk
Electricity System Development
Tonne Kjærsvej 65
7000 Fredericia, Denmark
e-mail: *ano@energinet.dk*

Vince Piacenti
Robert Bosch LLC
System Engineering, Diesel Fuel Systems
38000 Hills Tech Drive
Farmington Hills, MI 48331, USA
e-mail: *vince.piacenti@us.bosch.com*

Jörg Pieschel
Otto-von-Guericke University
Institute of Materials and Joining Technology
Universitätsplatz 2
39106 Magdeburg, Germany
e-mail: *pieschel@mb.uni-magdeburg.de*

Stefan Pischinger
RWTH Aachen University
Institute for Combustion Engines
Schinkelstr. 8
52062 Aachen, Germany
e-mail: *pischinger@vka.rwth-aachen.de*

Didier M. Priem
École Centrale Nantes
Department of Materials
1 Rue de la Noë, GEM UMR CNRS 6183
44321 Nantes, France
e-mail: *didier.priem@ec-nantes.fr*

Frank Riedel
Fraunhofer-Institute for Machine Tools and Forming Technology (IWU)
Department of Joining Technology
Reichenhainer Str. 88
09126 Chemnitz, Germany
e-mail: *frank.riedel@iwu.fraunhofer.de*

Holger Saage
University of Applied Sciences of Landshut
Faculty of Mechanical Engineering
Am Lurzenhof 1
84036 Landshut, Germany
e-mail: *holger.saage@fh-landshut.de*

Shuichi Sakamoto
Niigata University
Department of Mechanical and Production Engineering
Ikarashi 2-8050
950 2181 Niigata, Japan
e-mail: *sakamoto@eng.niigata-u.ac.jp*

Roger Schaufele
California State University
1250 Bellflower Boulevard
Long Beach, CA 90840, USA
e-mail: *rdschaufele@aol.com*

Markus Schleser
RWTH Aachen University
Welding and Joining Institute
Pontstr. 49
52062 Aachen, Germany
e-mail: *schleser@isf.rwth-aachen.de*

Meinhard T. Schobeiri
Texas A&M University
Department of Mechanical Engineering
College Station, TX 77843-3123, USA
e-mail: *tschobeiri@tamu.edu*

Miroslaw J. Skibniewski
University of Maryland
Department of Civil and Environmental Engineering
1188 Glenn L. Martin Hall
College Park, MD 20742-3021, USA
e-mail: *mirek@umd.edu*

Jagjit Singh Srai
University of Cambridge
Centre for International Manufacturing
Institute for Manufacturing
Cambridge, CB2 1 RX, UK
e-mail: *jss46@cam.ac.uk*

Vivek Srivastava
Corporate Technology Strategy Services
Aditya Birla Management Corporation
MIDC Taloja, Panvel
Navi Mumbai, India
e-mail: *vivek.srivastava@adityabirla.com*

Peter Stephan
Technical University Darmstadt
Institute of Technical Thermodynamics
Department of Mechanical Engineering
Petersenstr. 30
64287 Darmstadt, Germany
e-mail: *pstephan@ttd.tu-darmstadt.de*

Zbigniew A. Styczynski
Otto-von-Guericke University
Electric Power Networks and Renewable Energy Sources
Universitätsplatz 2
39106 Magdeburg, Germany
e-mail: *sty@ovgu.de or sty@ieee.org*

P.M.V. Subbarao
Indian Institute of Technology
Mechanical Engineering Department
HAUS KHAS
New Delhi, 110 016, India
e-mail: *pmvs@mech.iitd.ac.in*

Oliver Tegel
Dr.-Ing. h.c. F. Porsche AG
R&D, IS-Management
Porschestr.
71287 Weissach, Germany
e-mail: *oliver.tegel@porsche.de*

A. Erman Tekkaya
ATILIM University
Department of Manufacturing Engineering
Incek
Ankara, 06836, Turkey
e-mail: *etekkaya@atilim.edu.tr*

Klaus-Dieter Thoben
University of Bremen
Bremen Institute for Production and Logistics GmbH
Department of ICT Applications in Production
Hochschulring 20
28359 Bremen, Germany
e-mail: *tho@biba.uni-bremen.de*

Marcel Todtermuschke
Fraunhofer-Institute for Machine Tools and Forming Technology
Department of Assembling Techniques
Reichenhainer Str. 88
09126 Chemnitz, Germany
e-mail: *marcel.todtermuschke@saxonia.net*

Helmut Tschoeke
Otto-von-Guericke University
Institute of Mobile Systems
Universitätsplatz 2
39106 Magdeburg, Germany
e-mail: *tschoeke@ovgu.de*

Jon H. Van Gerpen
University of Idaho
Department of Biological and Agricultural Engineering
Moscow, ID, USA
e-mail: *jonvg@uidaho.edu*

Anatole Vereschaka
Moscow State University of Technology "STANKIN"
Department of Mechanical Engineering Technology and Institute of Design and Technological Informatics
Laboratory of Surface Nanosystems
Russian Academy of Science
Vadkovsky pereulok 1
Moscow, 101472, Russia
e-mail: *dr_averes@rambler.ru*

Detlef von Hofe
Hohen Dyk 106
47803 Krefeld, Germany
e-mail: *detlef.von.hofe@web.de*

Nikolaus Wagner
RWTH Aachen University
ISF Welding and Joining Institute
Pontstr. 49
52062 Aachen, Germany
e-mail: *wa@isf.rwth-aachen.de*

Jacek G. Wankowicz
Institute of Power Engineering
ul. Mory 8
01-330 Warsaw, Poland

Ulrich Wendt
Otto-von-Guericke University
Department of Materials and Joining Technology
Universitätsplatz 2
39106 Magdeburg, Germany
e-mail: *wendt@ovgu.de*

Steffen Wengler
Otto-von-Guericke University
Faculty of Mechanical Engineering
Institute of Manufacturing Technology
and Quality Management
Universitätsplatz 2
39106 Magdeburg, Germany
e-mail: *swengler@ovgu.de*

Bernd Wilhelm
Volkswagen AG
Sitech Sitztechnik GmbH
Stellfelder Str. 46
38442 Wolfsburg, Germany
e-mail: *bernd.wilhelm2@volkswagen.de*

Patrick M. Williams
Assystem UK
1 The Brooms, Emersons Green
Bristol, BS16 7FD, UK
e-mail: *pwilliams@assystemuk.com*

Lutz Wisweh
Otto-von-Guericke University
Faculty of Mechanical Engineering
Institute of Manufacturing Technology
and Quality Management
Universitätsplatz 2
39106 Magdeburg, Germany
e-mail: *lutz.wisweh@ovgu.de*

Johannes Wodara
Schweißtechnik-Consult
Hegelstr. 38
39104 Magdeburg, Germany
e-mail: *johanneswodara@hotmail.com*

Klaus Woeste
RWTH Aachen University
ISF Welding and Joining Institute
Pontstr. 49
52062 Aachen, Germany
e-mail: *wo@isf.rwth-aachen.de*

Hen-Geul Yeh
California State University
Department of Electrical Engineering
1250 Bellflower Boulevard
Long Beach, CA 90840-8303, USA
e-mail: *heyeh@csulb.edu*

Hsien-Yang Yeh
California State University Long Beach
Department of Mechanical and Aerospace
Engineering
1250 Bellflower Boulevard
Long Beach, CA 90840, USA
e-mail: *hyyeh@csulb.edu*

Shouwen Yu
Tsinghua University
School of Aerospace
Beijing, 100084, P.R. China
e-mail: *yusw@mail.tsinghua.edu.cn*

Contents

List of Abbreviations

3DP	3-D printing

A

ABCS	automated building construction systems
ABS	acrylonitrile-butadiene-styrene
ACCS	automatic cutter control system
ACFM	actual cubic feet per minute
ADAS	advanced driver-assistance system
ADI	austempered cast iron
ADI	austempered ductile cast iron
AFM	atomic force microscope
AGR	advanced gas-cooled reactor
API	application programming interface
ARIS	architecture of integrated information systems
AS	active sum
ASC	automatic stability control
ASME	American Society of Mechanical Engineers
ATC	automatic tool change
ATS	air transport system
ATZ	Automobiltechnische Zeitschrift
AWJ	abrasive waterjet

B

bcc	body-centered cubic
bct	body-centered tetragonal
BDC	bottom dead center
bdd	block definition diagram
BHN	Brinell hardness
BHS	Brinell hardness
BHW	Brinell hardness
BiW	body-in-white
BM	beam machining
BMEP	break mean effective pressure
BMS	bionic manufacturing system
BOM	bill of materials
BOO	bill of operations
BOSC	built-to-order supply chain
BPM	ballistic particle manufacturing
BPR	business process reengineering
BSE	backscattered electrons
BVP	boundary-value problem
BWB	blended wing body
BWR	boiling-water reactor

C

CAD	computer-aided design
CAES	compressed air energy storage
CAM	computer-aided manufacturing
CAM-LEM	computer-aided manufacturing of laminated engineering material
CAN	controller area network
CAPP	computer-aided process planning
CAS	computer-aided styling
CAS	calibrated airspeed
CBN	cubic boron nitride
CC	contour crafting
CCD	charge-coupled device
CCGT	combined cycle gas turbines
CCT	continuous cooling transition
ccw	counterclockwise
CD	compact disc
CD	continuous dressing
CDC	crank dead center
CDP	car development process
CDP	car development project
CE	concurrent engineering
CFC	chlorofluorocarbons
CFD	computational fluid dynamics
CFRP	carbon fiber reinforced plastic
CGI	compacted graphite iron
CHP	combined heat and power
CI	compression ignition
CI	corporate identity
CIFI	cylinder-individual fuel injection
CIM	computer-integrated manufacturing
CIMOSA	computer-integrated manufacturing open system architecture
CIP	continuous improvement process
CLFM	constitutional liquid film migration
CMCV	charge motion control valve
CMM	coordinate measuring machine
CMP	chemical-mechanical planarization
CMU	cooperative manufacturing unit
CNC	computer numerical control
CNG	compressed natural gas
CODAP	code francais de construction des appareils a pression
CPFR	collaborative planning, forecasting, and replenishment
CPM	critical-path method
CPT	critical pitting temperature
CR	common rail
CRM	customer relationship management
CRP	continuous replenishment planning

CRSS	critical resolved shear stress
CRT	cathode ray tube
CSLP	capacitated lot-sizing lead-time problem
CVD	chemical vapor deposition
CVN	charpy V-notch

D

DBTT	ductile to brittle transition
DC	direct current
DfC	design for construction
DFE	design for the environment
DFIG	double-fed induction generator
DfRC	design for robotic construction
DIC	differential interference contrast
DI	direct injection
DIN	Deutsches Institut für Normung
DIO	digital input output
DIS	Draft International Standard
DLF	direct laser fabrication
DLM	direct laser fabrication
DMD	direct metal deposition
DMLS	direct metal laser sintering
DMU	digital mock-up
DNC	direct numerical control
DPH	diamond-pyramid hardness number
DSC	differential scanning calorimetry
DVS	Verband für Schweißen und verwandte Verfahren e.V.
D/W	depth-to-width

E

E2	extended enterprises
EAS	equivalent airspeed
EBM	electron beam machining
EBSD	electron backscatter diffraction
ECDD	evanescent coupling display device
ECDM	electrochemical-discharge machining
ECG	electrochemical grinding
ECM	electrochemical machining
ECM	electronic control module
ECR	efficient customer response
ECU	electronic control unit
EDG	electro-discharge grinding
EDM	electro-discharge machining
EDM	engineering data management
EDP	electronic data processing
EDS	energy-dispersive x-ray spectroscopy
EDX	energy dispersive x-ray spectrometer
EELS	electron energy loss spectroscopy
EFFBD	enhanced functional flow block diagram
EGR	exhaust gas recirculation
EIS	entry into service
EJMA	Expansion Joint Manufacturer's Association
ELID	electrolytic in-process dressing
EMC	electromagnetic compatibility
EPA	Environmental Protection Agency
EPC	event-driven process chains
EP	extreme pressure
EPDM	ethylene propylene diene monomer
EPMA	electron probe microanalysis
ERP	enterprise resource planning
ESCA	electron spectroscopy for chemical analysis
ESP	electrostatic precipitator
ESP	electronic stability program

F

FAR	federal air regulations
FBC	fluidized-bed combustion
FBR	fast breeder reactor
fcc	face-centered cubic
FD	forced draught
FDM	fused deposition modeling
FE	flap-extended
FEGT	furnace exit gas temperature
FEM	finite element modeling
FEPA	Federation of European Producers of Abrasíves
FFT	fast Fourier transform
FGD	flue gas desulphurization
FKA	Forschunggesellschaft Kraftfahrwesen mbH Aachen
FIB	focused ion beam
FLD	forming limit diagram
FMEA	failure mode and effect analysis
FPM	freeform powder molding
FPO	future project office

G

GA	general arrangement
GERAM	generalized enterprise reference model architecture and methodology
GHG	greenhouse gas
GIM	GRAI integrated methodology
GJL	lamellar graphite cast iron
GMA	gas metal arc
GoM	guidelines of modeling
GPS	global positioning system
G/R	gradient/growth rate
GTAW	gas tungsten arc welding

H

HAZ	heat-affected zone
HC	hydrocarbons
HCP	hexagonal closed packed
hcp	hexagonal closed packed

HDC	head dead center
HDPE	high-density polyethylene
HEM	high-efficiency machining
HFID	heated flame ionization detector
HHV	higher heating value
HIL	hardware-in-the-loop
HIP	hot isostatic pressing
HMS	holonic manufacturing systems
HP	high pressure
HPCC	high-pressure combustion chamber
HPT	high-pressure turbine
HRC	Rockwell hardness
HRSG	heat recovery steam generator
HSC	high-speed cutting
HSLA	high-strength low-alloy
HSM	high-speed machining
HSS	high-speed steel
HTA	heavier than air
HVDC	high-voltage direct-current

I

IAARC	International Association for Automation and Robotics in Construction
IAS	indicated airspeed
IBD	internal block diagram
IBM	ion beam machining
ICAO	International Civil Aviation Organization
ICDD	International Center for Diffraction Data
ICE	internal combustion engines
ICE	intercity express
IC	integrated circuits
ICT	information and communication technology
IDD	interferometric display device
IDI	indirect diesel injection
ID	induced draught
ID	inside diameter
IEEE	Institute of Electrical and Electronics Engineers
IE	Erichson index
IFAC	International Federation for Automatic Control
IFIP	International Federation for Information Processing
IGBT	insulated gate bipolar transistor
IGC	intergranular corrosion test
IGES	initial graphics exchange specification
IIE	information-interoperable environment
IISE	ion-induced secondary electrons
ILT	Fraunhofer Institut für Lasertechnik
IMP	International Marketing and Purchasing
IP	intermediate pressure
ISB	interact system B
ISARC	International Symposia on Automation and Robotics in Construction
ISO	International Standards Organization
IT	information technology
IVP	initial-value problem

J

JIT	Java intelligent network
JiT	just-in-time

L

LAM	laser-assisted machining
LB	laser beam
LBM	laser beam machining
LCA	life cycle analysis
LCI	life cycle inventory
LC	laser cutting
LDV	light duty vehicles
LENS	laser engineered net shaping
LHV	lower heating value
LMJ	micro-jet procedure
LM	layer manufacturing
LNG	liquefied natural gas
LOM	laminated object manufacturing
LP	low pressure
LPCC	low-pressure combustion chamber
LPG	petroleum gas
LPT	low-pressure turbine
LRO	long-range order
LTA	lighter than air
LYS	lower yield stress

M

MAM	motorized air cycle machine
MAP	main air pipe
MAS	multi-agent systems
MCD	monocrystalline diamond
MDT	mean down time
MEMS	microelectromechanical system
MEP	mean effective pressure
MESA	Manufacturing Enterprise Solutions Association
MES	manufacturing execution systems
MHD	magnetohydrodynamics
MIC	microbiologically influenced corrosion
MIPS	microprocessor without interlocked pipeline stages
MLW	maximum landing weight
MMC	metal-matrix composites
MOSFET	metal oxide semiconductor field effect transistor
MPI	magnetic particle inspection
MPM	metra potential method
MPW	magnetic pulse welding
MRI	magnetic resonance imaging

MRP	manufacturing resources planning
MRP	materials requirement planning
M/T	machine tool
MTBE	methyl *t*-butyl ether
MTBF	mean time between failure
MWE	manufacturers weight empty
MZFW	maximum zero fuel weight

N

NACE	National Association of Corrosion Engineers
NC	numerically controlled
NCE	numerically controlled equipment
NDE	nondestructive evaluation
NDI	nondestructive inspection
NDIR	nondispersive infrared
ND	normal direction
NDT	nondestructive testing
NEDC	New European Driving Cycle
NEMS	nanoelectromechanical systems
NLGI	National Association of Lubricating Grease Institute
NTP	normal temperature and pressure
NV-EBW	nonvacuum electron-beam welding
NVH	noise–vibration–harshness

O

OBJ	polygon mesh
ODE	ordinary differential equation
OECD	Organisation for Economic Co-operation and Development
OFA	over fire air
OFW	oblique flying wing
OIM	orientation imaging microscopy
OLE	object linking and embedding
OMT	object-modeling technique
OOSE	object-orientes software engineering
OPC	open connectivity via open standards
ORiN	open robot interface for the network
OWE	operating weight empty

P

PABADIS	plant automation based on distributed systems
PAM	plasma arc machining
PBM	plasma beam machining
PBMR	pebble-bed reactor
PC	pulverized coal
PC	polycrystalline
PC	personal computer
PCBN	polycrystalline cubic boron nitride
PCD	polycrystalline diamond
PCM	powertrain control module
PDE	partial differential equations
PDF	powder diffraction file
PDM	product data management
PEMFC	polymer electrolyte fuel cell
PERA	purdue enterprise reference architecture
PERT	project evaluation and review technique
PET	polyethylene terephthalate
PHE	plate heat exchanger
PLC	programmable logic controller
PLS	pre-lining support
PM	powder metallurgy
PMZ	partially melted zone
PPC	production planning and control
ppm	parts per million
PQR	procedure qualification record
PROSA	product–resource–order–staff architecture
PSB	persistent slip bands
PSD	power spectral densities
PSLX	planning and scheduling language on XML specifications
PS	passive sum
p.t.o.	power take-off
PVC	polyvinyl chloride
PVD	physical vapor deposition
PV	pressure valve
PWB	printed wiring board
PWHT	post-weld heat treatment
PWR	pressurized-water reactor

Q

QA	quality assurance
QCC	quality control charts
QFD	quality function deployment
QMS	quality management systems

R

RAC	robot action command
RAMS	reliability, availability, maintainability, safety
RAO	robot access object
RaoSQL	robot access object SQL
RAP	reclaimed asphalt pavements
RBV	resource-based view
RD	rolling direction
RE	reverse engineering
RF	radiofrequency
RFID	radiofrequency identification
RIE	reactive ion etching
RISC	reduced-instruction-set computer
RK	Runge–Kutta method
RM	rapid manufacturing
RP	rapid prototyping
RPI	Rensselaer Polytechnic Institute
rpm	revolutions per minute

RPZ	risk priority number
RRD	robot resource definition
RT	radiographic testing
RT	reheat turbine
RTM	resin transfer molding
RT	room temperature
rms	root mean square
RUP	rational unified process

S

SAES	scanning Auger electron spectroscopy
SBR	polystyrene-butadien-rubber
SC	supply chain
SC	supercritical
SCADA	supervisory control and data aquisition
SCF	steel-frame buildings
SCF	super construction factory
SCM	supply chain management
SCOR	supply-chain operations reference
SC	supply chain
SCTR	solidification cracking temperature range
SDM	shape deposition manufacturing
SEDM	spark electro-discharge machining
SEFI	sequential fuel injection
SEM	scanning electron microscopy
SE	secondary electrons
SFC	specific fuel consumption
SGC	solid ground curing
SHE	standard hydrogen electrode
SHM	structural health monitoring
SI	spark ignition
SI	secondary ions
SI	spark-ignited
SI	system international
SIC	statistical inventory control
SIMS	secondary-ion mass spectroscopy
SLA	stereolithography
SLCA	streamlined life cycle analysis
SLPL	space limit payload
SLS	selective laser sintering
SMART	Shimizu manufacturing system by advanced robotics technology
SMAW	shielded metal arc welding
SMD	surface mounted device
SME	small and medium-sized enterprises
SMM	Sanders model maker
SNCR	selective noncatalytic reduction systems
SNG	synthetic natural gas
SN	supply network
SoA	space of activity
SOF	soluble organic fraction
SOHC	single overhead camshaft
SOP	start of production
SPC	statistical process control
SPV	simple pressure vessel
SQL	structured query language
SRO	short-range order
STL	stereolithography language
SUV	sports utility vehicle
SysML	systems modelling language

T

TCL	total accumulated crack length
TCT	time compression technology
TDC	top dead center
TD	transversal direction
TEMA	Tubular Exchanger Manufacturer's Association
TEM	transmission electron microscopy
TGV	train à grande vitesse
TIG	gas tungsten arc welding
TLAR	top-level aircraft requirements
TMAH	tetramethyl ammonium hydroxide
TMC	traffic message channel
TOR	top of rail
TPM	total productive maintenance
TPS	Toyota production system
TQM	total quality management
TRIAC	triode alternating current switch
TSF	topographic shell fabrication
TTS	tribotechnical system
TTT	time–temperature transition

U

UCAV	unmanned combat air vehicle
UHC	unburned hydrocarbon
UHCA	ultra-high-capacity aircraft
UHEGT	ultra high efficiency gas turbine technology
UIC	Union International des Chemins de Fer
ULEV	ultralow-emission vehicle
UNS	unified numbering system
UPS	uninterruptible power supply
UPV	unifired pressure vessel
US	ultrasonic
USC	ultra-supercritical steam
USM	ultrasonic machining
UTS	ultimate tensile strength
UT	ultrasonic testing
UYS	upper yield stress

V

VC	vacuum casting
VDI	Verein Deutscher Ingenieure (Association of German Engineers)
VHN	Vickers hardness number
VICS	Voluntary Interindustry Commerce Standard Association

VI	viscosity index
VLCT	very large commercial transport
VOC	volatile organic compound
VOF	volatile organic fraction
VO	virtual organizations
VPN	virtual private network
VR	virtual-reality
VTOL	vertical take-off and landing

W

WBS	work breakdown structure
WDS	wavelength dispersive x-ray spectroscopy
WDX	wavelength dispersive x-ray spectroscopy
WEDM	wire electro-discharge machining
WLT	white light triangulation
WPS	weld procedure specification
WSP	wheel-slide protection
WWW	world wide web
W/C	water/cement

X

XPS	x-ray-exited photoelectron spectroscopy
XRD	x-ray diffraction

Y

YPE	yield point elongation

Z

ZEV	zero-emission vehicle

Part A Fundamentals of Mechanical Engineering

Introduction

1. Introduction to Mathematics for Mechanical Engineering

Ramin S. Esfandiari

This chapter is concerned with fundamental mathematical concepts and methods pertaining to mechanical engineering. The topics covered include complex analysis, differential equations, Laplace transformation, Fourier analysis, and linear algebra. These basic concepts essentially act as tools that facilitate the understanding of various ideas, and implementation of many techniques, involved in different branches of mechanical engineering. Complex analysis, which refers to the study of complex numbers, variables and functions, plays an important role in a wide range of areas from frequency response to potential theory. The significance of ordinary differential equations (ODEs) is observed in situations involving the rate of change of a quantity with respect to another. A particular area that requires a thorough knowledge of ODEs is the modeling, analysis, and control of dynamic systems. Partial differential equations (PDEs) arise when dealing with quantities that are functions of two or more variables; for instance, equations of motions of beams and plates. Higher-order differential equations are generally difficult to solve. To that end, the Laplace transformation is used to transform the data from the time domain to the so-called *s*-domain, where equations are algebraic and hence easy to treat. The solution of the differential equation is ultimately obtained when information is transformed back to time domain. Fourier analysis is comprised of Fourier series and Fourier transformation. Fourier series are a specific trigonometric series representation of a periodic signal, and frequently arise in areas such as system response analysis. Fourier transformation maps information from the time to the frequency domain, and its extension leads to the Laplace transformation. Linear algebra refers to the study of vectors and matrices, and plays a central role in the analysis of systems with large numbers of degrees of freedom.

1.1 Complex Analysis

Complex numbers, variables and functions are the main focus of this section. We will begin with complex numbers, their representations, as well as properties. The idea is then extended to complex variables and their functions.

1.1.1 Complex Numbers

A complex number z appears in the rectangular form

$$z = x + \mathrm{i}y\,,$$
$$\mathrm{i} = \sqrt{-1} = \text{imaginary number}\,, \tag{1.1}$$

where x and y are real numbers, called the real and imaginary parts of z, respectively, and denoted by $x = \mathrm{Re}(z)$, $y = \mathrm{Im}(z)$. For example, if $z = -1 + 2\mathrm{i}$, then

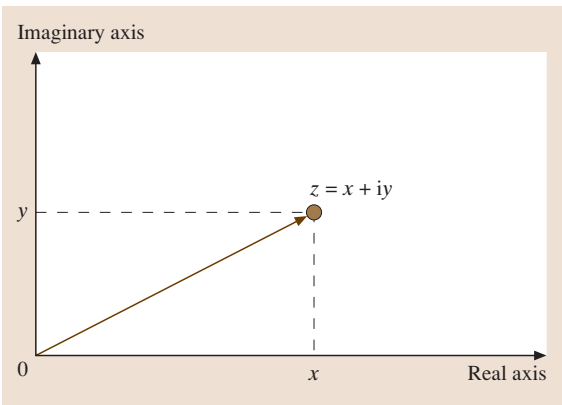


Fig. 1.1 Geometrical representation of complex numbers – the complex plane

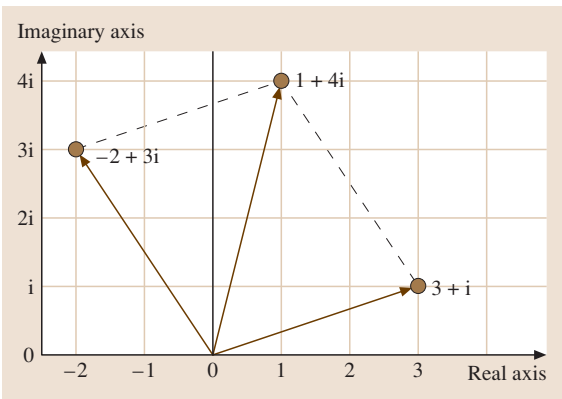


Fig. 1.2 Addition of complex numbers by vector addition

$x = \mathrm{Re}(z) = -1$ and $y = \mathrm{Im}(z) = 2$. A complex number with zero real part is known as pure imaginary, e.g., $z = 4\mathrm{i}$. Two complex numbers are said to be equal if and only if their respective real and imaginary parts are equal. Addition of complex numbers is performed component-wise, that is, if $z_1 = x_1 + \mathrm{i}y_1$ and $z_2 = x_2 + \mathrm{i}y_2$, then

$$\begin{aligned} z_1 + z_2 &= (x_1 + \mathrm{i}y_1) + (x_2 + \mathrm{i}y_2) \\ &= (x_1 + x_2) + \mathrm{i}(y_1 + y_2)\,. \end{aligned} \tag{1.2}$$

Multiplication of two complex numbers is performed in the same way as two binomials with the provision that $\mathrm{i}^2 = -1$, $\mathrm{i}^3 = -\mathrm{i}$, $\mathrm{i}^4 = 1$, etc. need be taken into account, that is,

$$\begin{aligned} z_1 z_2 &= (x_1 + \mathrm{i}y_1)(x_2 + \mathrm{i}y_2) \\ &= x_1 x_2 + \mathrm{i}y_1 x_2 + \mathrm{i}x_1 y_2 + \mathrm{i}^2 y_1 y_2 \\ &= (x_1 x_2 - y_1 y_2) + \mathrm{i}(x_1 y_2 + x_2 y_1)\,. \end{aligned} \tag{1.3}$$

Complex Plane

Since complex numbers consist of a real part and an imaginary part, they have a two-dimensional character, and hence may be represented geometrically as points in a Cartesian coordinate system, known as the complex plane. The x-axis of the complex plane is the real axis, and its y-axis is called the imaginary axis, (Fig. 1.1). Noting that $z = x + \mathrm{i}y$ is uniquely identified by an ordered pair (x, y) of real numbers, we can represent z as a two-dimensional (2-D) vector in the complex plane, with initial point 0 and terminal point $z = x + \mathrm{i}y$; in other words, the position vector of the point z. The imaginary number i, for instance, can be identified by (0, 1). So, the concept of vector addition also applies to the addition of complex numbers. For that, let us consider $z_1 = -2 + 3\mathrm{i}$ and $z_2 = 3 + \mathrm{i}$ in Fig. 1.2. It is then evident that their sum, $z_1 + z_2 = 1 + 4\mathrm{i}$, is exactly what we would obtain by adding the corresponding position vectors of z_1 and z_2.

The magnitude of a complex number $z = x + \mathrm{i}y$ is defined as

$$|z| = \sqrt{x^2 + y^2}\,. \tag{1.4}$$

Geometrically, $|z|$ is the distance from z to the origin of the complex plane. If z is real, it must be located on the x-axis, and its magnitude is equal to its absolute value. If z is pure imaginary ($z = \mathrm{i}y$), then it is on the y-axis, and $|z| = |y|$. The quantity $|z_1 - z_2|$ gives the distance between z_1 and z_2 (Fig. 1.3).

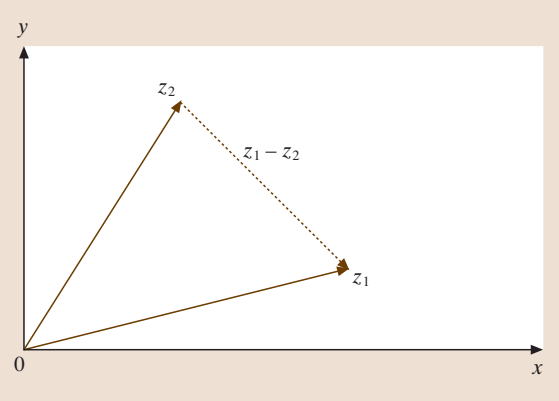


Fig. 1.3 Distance between two complex numbers

Example 1.1: Distance
Given $z_1 = 5+2\mathrm{i}$ and $z_2 = -1+10\mathrm{i}$, then

$$|z_1 - z_2| = |(5+2\mathrm{i}) - (-1+10\mathrm{i})| = |6-8\mathrm{i}| = \sqrt{6^2+(-8)^2} = 10$$

Addition of complex numbers obeys the triangle inequality,

$$|z_1 + z_2| \le |z_1| + |z_2| \,. \tag{1.5}$$

Given a complex number $z = x+\mathrm{i}y$, its conjugate, denoted by $\bar{z}$, is defined as $\bar{z} = x-\mathrm{i}y$. An immediate result is that the product of a complex number ($z \neq 0$) and its conjugate is a positive, real number, equal to the square of its magnitude, that is,

$$z\bar{z} = (x+\mathrm{i}y)(x-\mathrm{i}y) = x^2+y^2 \,. \tag{1.6}$$

Geometrically, a complex number and its conjugate are reflections of one another about the real axis; (Fig. 1.4).

Example 1.2: Conjugation
Given $z = -1+2\mathrm{i}$, we have

$$z\bar{z} = (-1+2\mathrm{i})\overline{(-1+2\mathrm{i})} = (-1+2\mathrm{i})(-1-2\mathrm{i}) = (-1)^2+(2)^2 = 5 \,,$$

which agrees with $|z| = |-1+2\mathrm{i}| = \sqrt{5}$.

Complex conjugation is extremely useful in complex algebra. To begin with, noting that

$$z+\bar{z} = (x+\mathrm{i}y)+(x-\mathrm{i}y) = 2x$$

and

$$z-\bar{z} = (x+\mathrm{i}y)-(x-\mathrm{i}y) = 2\mathrm{i}y$$

we conclude that

$$x = \mathrm{Re}(z) = \frac{1}{2}(z+\bar{z}) \,,$$
$$y = \mathrm{Im}(z) = \frac{1}{2\mathrm{i}}(z-\bar{z}) \,. \tag{1.7}$$

Division of Complex Numbers

Let us consider z_1/z_2 where $z_1 = x_1+\mathrm{i}y_1$ and $z_2 = x_2+\mathrm{i}y_2 (\neq 0)$. Multiply the numerator and the denominator by the conjugate of the denominator, that is, $\bar{z}_2 = x_2 - \mathrm{i}y_2$. Then, by (1.6), the resulting denominator is simply $|z_2|^2 = x_2^2+y_2^2$, a real number. In summary,

$$\begin{aligned}\frac{x_1+\mathrm{i}y_1}{x_2+\mathrm{i}y_2} &= \frac{x_1+\mathrm{i}y_1}{x_2+\mathrm{i}y_2}\frac{x_2-\mathrm{i}y_2}{x_2-\mathrm{i}y_2} \\ &= \frac{(x_1x_2+y_1y_2)+\mathrm{i}\,(y_1x_2-y_2x_1)}{x_2^2+y_2^2} \\ &= \frac{x_1x_2+y_1y_2}{x_2^2+y_2^2} + \frac{y_1x_2-y_2x}{x_2^2+y_2^2}\mathrm{i} \end{aligned} \tag{1.8}$$

where the outcome is represented in the standard rectangular form.

Example 1.3: Division of complex numbers
Perform the following division of complex numbers, and express the result in the standard rectangular form:

$$\frac{2-\mathrm{i}}{-1+4\mathrm{i}}$$

Solution. Multiplication and division by the conjugate of the denominator, yields

$$\begin{aligned}\frac{(2-\mathrm{i})(-1-4\mathrm{i})}{(-1+4\mathrm{i})(-1-4\mathrm{i})} &= \frac{-6-7\mathrm{i}}{17} \\ &= -\frac{6}{17} - \mathrm{i}\frac{7}{17}\end{aligned}$$

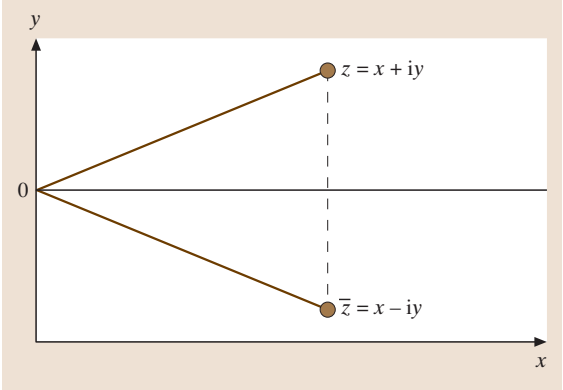


Fig. 1.4 A complex number and its conjugate

Polar Representation of Complex Numbers

Although the standard rectangular form is suitable in certain instances, it is quite inconvenient in most others. For example, imagine the simplification of $(-2+3\mathrm{i})^{10}$. Situations of this type require a special form that simplifies the complex algebra. The polar form of a complex number, as suggested by its name, uses the polar coordinates to represent a complex number in the complex plane. Recall that any point in the plane can be determined by a radial coordinate r and an angular coordinate θ. So, the same holds for a complex number $z = x + \mathrm{i}y \neq 0$ in the complex plane, (Fig. 1.5). The relationship between the rectangular and polar coordinates is given by

$$x = r\cos\theta\,, \quad y = r\sin\theta\,. \tag{1.9}$$

We first introduce Euler's formula,

$$\mathrm{e}^{\mathrm{i}\theta} = \cos\theta + \mathrm{i}\sin\theta\,. \tag{1.10}$$

Then, (1.9) and (1.10) yield

$$z = x + \mathrm{i}y = r\cos\theta + \mathrm{i}\,(r\sin\theta) = r\,\mathrm{e}^{\mathrm{i}\theta}$$

In summary,

$$z = r\,\mathrm{e}^{\mathrm{i}\theta}\,, \tag{1.11}$$

which is called the polar form of the complex number z. Here, the magnitude (or modulus) of z is defined by

$$r = |z| = \sqrt{x^2 + y^2} = \sqrt{z\bar{z}} \tag{1.12}$$

and the phase (or argument) of z is

$$\theta = \arg z = \tan^{-1}\left(\frac{\mathrm{Im}(z)}{\mathrm{Re}(z)}\right) = \tan^{-1}\left(\frac{y}{x}\right)\,. \tag{1.13}$$

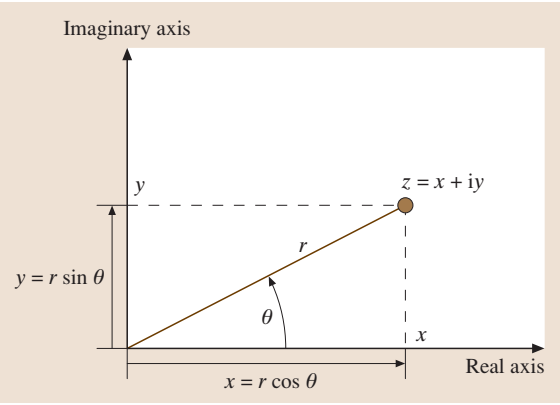


Fig. 1.5 Relation between the rectangular and polar forms of a complex number

The angle θ is measured from the positive real axis and, by convention, is regarded as positive in the sense of the counterclockwise (ccw) direction. It is measured in radians (rad) and is determined in terms of integer multiples of 2π. The specific value of θ that lies in the interval $(-\pi, \pi]$ is called the principal value of $\arg z$ and is denoted by $\arg z$. In engineering analysis, it is also common to express the polar form of z as

$$z = r\angle\theta \tag{1.14}$$

where $\angle$ denotes the angle.

Example 1.4: Phase via location
Express $z = \frac{2}{-1+\mathrm{i}}$ in polar form.

Solution. First, express z in standard rectangular form, as

$$z = \frac{2}{-1+\mathrm{i}}\frac{-1-\mathrm{i}}{-1-\mathrm{i}} = \frac{-2-2\mathrm{i}}{2} = -1-\mathrm{i}\,,$$

indicating that z is located in the third quadrant of the complex plane. Next, we use (1.13) to find

$$\theta = \tan^{-1}\left(\frac{-1}{-1}\right) = 45^\circ = \frac{\pi}{4}\mathrm{rad}\,.$$

However, the only information this provides is that the (smallest) angle between OA and the real axis (Fig. 1.6) is 45°. Since z is in the third quadrant, its actual phase is then $180 + 45 = 225^\circ$ ($\pi + \pi/4 = 5\pi/4$ rad) if measured in the ccw direction, or -135° ($-3\pi/4$ rad) in the clockwise (cw) direction. So, the polar form of z can be written as

$$z = -1-\mathrm{i} = \sqrt{2}\,\mathrm{e}^{\mathrm{i}(5\pi/4)} \quad \text{or} \quad z = \sqrt{2}\angle\frac{5\pi}{4}\,.$$

Multiplication and Division in Polar Form. As cited earlier, polar form substantially reduces complex alge-

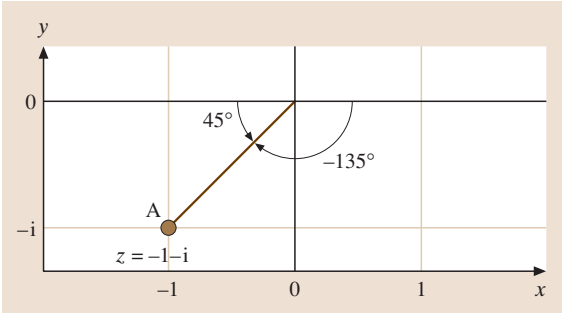


Fig. 1.6 Example 1.4

bra, in particular, multiplication and division. Consider two complex numbers $z_1 = r_1 e^{i\theta_1}$ and $z_2 = r_2 e^{i\theta_2}$. Subsequently,

$$z_1 z_2 = r_1 r_2 e^{i(\theta_1+\theta_2)} \quad \text{or} \quad r_1 r_2 \angle(\theta_1+\theta_2) . \tag{1.15}$$

This means the magnitude and phase of the product $z_1 z_2$ are

$$\begin{aligned} |z_1 z_2| &= r_1 r_2 = |z_1||z_2| \quad \text{and} \\ \arg(z_1 z_2) &= \theta_1 + \theta_2 = \arg(z_1) + \arg(z_2) \end{aligned} \tag{1.16}$$

Similarly, for division of complex numbers, we have

$$\frac{z_1}{z_2} = \frac{r_1}{r_2} e^{i(\theta_1-\theta_2)} \quad \text{or} \quad \frac{r_1}{r_2} \angle(\theta_1-\theta_2) . \tag{1.17}$$

so that

$$\begin{aligned} \left|\frac{z_1}{z_2}\right| &= \frac{r_1}{r_2} = \frac{|z_1|}{|z_2|} \quad \text{and} \\ \arg\left(\frac{z_1}{z_2}\right) &= \theta_1 - \theta_2 = \arg(z_1) - \arg(z_2) \end{aligned} \tag{1.18}$$

Complex Conjugation in Polar Form. Given the polar form of a complex number, $z = r e^{i\theta}$, its conjugate is obtained as

$$\begin{aligned} \bar{z} &= x - iy = r\cos\theta - i(r\sin\theta) \\ &= r(\cos\theta - i\sin\theta) \\ &= r[\cos(-\theta) + i\sin(-\theta)] \overset{\text{Euler's formula}}{=} r e^{-i\theta} . \end{aligned} \tag{1.19}$$

This result makes sense geometrically, since a complex number and its conjugate are reflections of one another through the real axis. Hence, they are equidistant from the origin, that is, $|z| = |\bar{z}| = r$, and the phase of one is the negative of the phase of the other, i. e., $\arg(z) = -\arg(\bar{z})$; Fig. 1.7. The important property of complex conjugation (1.6) can now be confirmed in polar form, as

$$z\bar{z} = (r e^{i\theta})(r e^{-i\theta}) = r^2 = |z|^2 .$$

Integer Powers of a Complex Number

The effectiveness of the polar form may further be demonstrated when raising a complex number to an *integer* power. Letting $z = r e^{i\theta}$, then

$$\begin{aligned} z^n &= (r e^{i\theta})^n = r^n e^{in\theta} \\ &\overset{\text{Euler's formula}}{=} r^n(\cos n\theta + i\sin n\theta) , \end{aligned} \tag{1.20}$$

so that $\mathrm{Re}(z^n) = r^n \cos n\theta$ and $\mathrm{Im}(z^n) = r^n \sin n\theta$.

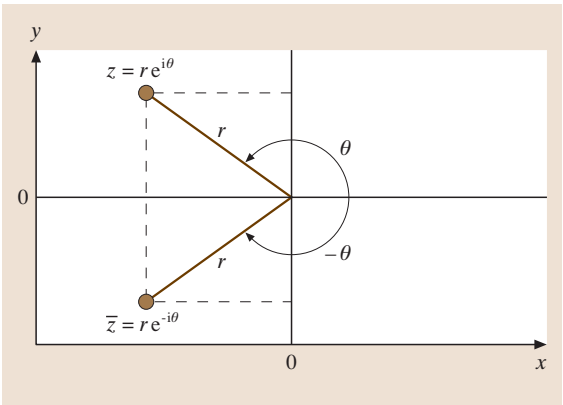


Fig. 1.7 A complex number and its conjugate in polar form

Roots of a Complex Number

In real calculus, if a is a real number then $\sqrt[n]{a}$ has a single value. On the contrary, given a complex number $z \neq 0$, and a positive integer n, then the nth root of z, written $\sqrt[n]{z}$, is *multivalued*. In fact, there are n different values of $\sqrt[n]{z}$, corresponding to each value of $z \neq 0$. For a known $z = r e^{i\theta}$, it can be shown that [1.1, 2]

$$\begin{aligned} \sqrt[n]{z} &= \sqrt[n]{r}\left(\cos\frac{\theta+2k\pi}{n} + i\sin\frac{\theta+2k\pi}{n}\right) , \\ k &= 0, 1, \cdots, n-1 . \end{aligned} \tag{1.21}$$

Geometrically, these n values are described as follows:

1. they all lie on a circle centered at the origin with a radius of $\sqrt[n]{r}$, and
2. they are the n vertices of an n-sided regular polygon.

Example 1.5: Fourth roots of unity
We are seeking $\sqrt[4]{z}$, where $z = 1$. Noting that $z = 1$ is on the positive real axis, one unit from the origin, we conclude that $r = 1$ and $\theta = 0$, hence $z = 1 = 1 e^{i(0)}$. Following (1.21), we find the four roots to be 1, i, -1, and $-$i; Fig. 1.8. Note that all four roots lie on a circle of radius $\sqrt[4]{1} = 1$ centered at the origin (the so-called unit circle), and are the vertices of a regular four-sided polygon, as asserted.

1.1.2 Complex Variables and Functions

If x or y or both vary, then $z = x + iy$ is referred to as a complex variable. The most well-known complex variable is the *Laplace variable* (Sect. 1.3). Letting S be a set of complex numbers, a function f defined on S is a *rule*, which assigns a complex number w to each

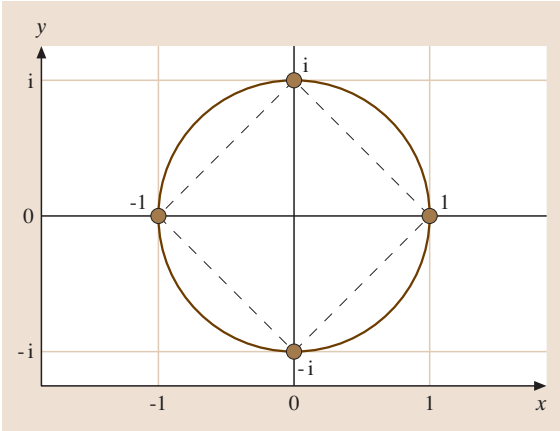


Fig. 1.8 Locations of the fourth roots of unity

$z \in S$. The notation is $w = f(z)$ and the set S is called the domain of definition of f. As an example, the domain of the function $w = z/(3-z)$ is any region that does not contain the point $z = 3$. Because z assumes different values from S, it is clearly a complex variable. Since w is complex, it must have a real part u and an imaginary part v, or $w = u + \mathrm{i}v$. Also $w = f(z)$ implies that w is dependent on $z = x + \mathrm{i}y$. Therefore, w depends on x and y, which means u and v depend on x and y, or

$$w = f(z) = u(x, y) + \mathrm{i}v(x, y) \,. \tag{1.22}$$

In real calculus, much can be learned about a function through its graph. However, when z and w are both complex, such a convenient graph of $w = f(z)$ is no longer available. This is because each z and w is located in a plane rather than on a line; more exactly, each z is in the xy-plane and each w in the uv-plane. However, a function f can still be thought of as a mapping (or transformation) that defines correspondence between points $z = (x, y)$ and $w = (u, v)$. Then, the *image* of a point $z \in S$ is the point $w = f(z)$, and the set of images of all points $z \in S$ that are mapped by f is the range of f, and denoted by

$$\{w | w = f(z), z \in S\} \,.$$

Analytic Functions

A function that is differentiable only at a single point is not of practical interest to us. What is of interest, however, is a function that is differentiable at a point and an entire neighborhood of that point. A neighborhood of a point is an open circular disk centered at the point. A complex function $f(z)$ is analytic (or holomorphic) at a point z_0 if it is differentiable throughout *some* neighborhood of z_0. A function is *analytic in a domain* if it is analytic at all points of that domain. Analytic functions arise in such areas as fluid flow and complex potentials.

Test of Analyticity: Cauchy–Riemann Equations. Suppose $f(z) = u(x, y) + \mathrm{i}v(x, y)$ is defined and continuous in a neighborhood of some point $z = x + \mathrm{i}y$, and that $f'(z)$ exists at that point. Then, the first partial derivatives of u and v with respect to x and y (that is, u_x, u_y, v_x, v_y) exist at that point and satisfy the Cauchy–Riemann equations

$$u_x = v_y, \quad v_x = -u_y \,. \tag{1.23}$$

Consequently, if $f(z)$ is analytic in some domain R, then the Cauchy–Riemann equations hold at every point of R.

Example 1.6: Cauchy–Riemann equations
Decide whether $f(z) = (\mathrm{i}-2)z^2 - 2\mathrm{i}z + \mathrm{i}$ is analytic.

Solution. Inserting $z = x + \mathrm{i}y$, we find

$$u(x, y) = -2xy - 2x^2 + 2y^2 + 2y \,,$$
$$v(x, y) = x^2 - y^2 - 4xy - 2x + 1 \,.$$

Partial differentiation yields

$$u_x = -2y - 4x \,, \quad v_y = -2y - 4x \,,$$
$$v_x = 2x - 4y - 2 \,, \quad u_y = -2x + 4y + 2 \,,$$

so that the Cauchy–Riemann equations hold for all z, and $f(z)$ is analytic for all z.

Cauchy–Riemann Equations in Polar Form. In many cases it is advantageous to use the polar form of the Cauchy–Riemann equations to test the analyticity of a function. The idea is to express z in polar form

$$z = r\,\mathrm{e}^{\mathrm{i}\theta} = r(\cos\theta + \mathrm{i}\sin\theta)$$

so that the function $f(z) = u(x, y) + \mathrm{i}v(x, y)$ can be written as $f(z) = u(r, \theta) + \mathrm{i}v(r, \theta)$. In this event, the Cauchy–Riemann equations in polar form can be derived as [1.1]

$$u_r = \frac{1}{r} v_\theta, \quad v_r = -\frac{1}{r} u_\theta \,. \tag{1.24}$$

This is particularly useful when dealing with z^m for $m \geq 3$, making it much easier to work with than $(x + \mathrm{i}y)^m$.

1.2 Differential Equations

Mathematical models of dynamic systems – mechanical, electrical, electromechanical, liquid-level, etc. – are represented by differential equations [1.3]. Therefore, it is imperative to have a thorough knowledge of their basic properties and solution techniques. In this section we will discuss the fundamentals of differential equations, specifically, ordinary differential equations (ODEs), and present analytical and numerical methods to solve them. Differential equations are divided into two general categories: *ordinary differential equations* and *partial differential equations (PDEs)*. An equation involving an unknown function and one or more of its derivatives is called a differential equation. When there is only one independent variable, the equation is called an ordinary differential equation (ODE). For example, $y' + 2y = e^x$ is an ODE involving the unknown function $y(x)$, its first derivative $y' = dy/dx$, as well as a given function e^x. Similarly, $xy'' - yy' = \sin x$ is an ODE relating $y(x)$ and its first and second derivatives with respect to x, as well as the function $\sin x$. While dealing with time-varying functions – as in many physical applications – the independent variable x will be replaced by t, representing time. In that case, the rate of change of the quantity $y = y(t)$ with respect to the independent variable t is denoted by $\dot{y} = dy/dt$. If the unknown function is a function of more than one independent variable, e.g., $u(x, y)$, the equation is referred to as a partial differential equation. The derivative of the highest order of the unknown function $y(x)$ with respect to x is the order of the ODE; for instance, $y' + 2y = e^x$ is of order one and $xy'' - yy' = \sin x$ is of order two. Consider an nth-order ordinary differential equation in the form

$$a_n y^{(n)} + a_{n-1} y^{(n-1)} + \cdots + a_1 y' + a_0 y = g(x) , \tag{1.25}$$

where $y = y(x)$ and $y^{(n)} = d^n y/dx^n$. If all coefficients $a_0, a_1, \cdots, a_n$ are either constants or functions of the independent variable x, then the ODE is linear. Otherwise, the ODE is nonlinear. Based on this, $y' + 2y = e^x$ describes a linear ODE, while $xy'' - yy' = \sin x$ is nonlinear.

Example 1.7: Order and linearity
Consider $3y''' - (2x+1)y'' + y = e^x$. Since the derivative of the highest order is three, the ODE is third order. Comparison with (1.25) reveals that $n = 3$, and $a_3 = 3$, $a_2 = -(2x+1)$, $a_1 = 0$, $a_0 = 1$, and $g(x) = e^x$. Thus, the ODE is linear.

1.2.1 First-Order Ordinary Differential Equations

First-order ODEs generally appear in the implicit form

$$F(x, y, y') = 0 . \tag{1.26}$$

For example, $y' + y^2 = \cos x$ can be expressed in the above form with $F(x, y, y') = y' + y^2 - \cos x$. In other cases, the equation may be written explicitly as

$$y' = f(x, y) . \tag{1.27}$$

An example would be $y' + 2y = e^x$ where $f(x, y) = e^x - 2y$. A function $y = s(x)$ is a solution of the first-order ODE in (1.26) on a specified (open) interval if it has a derivative $y' = s'(x)$ and satisfies (1.26) for all values of x in the given interval. If the solution is in the form $y = s(x)$, then it is called an explicit solution. Otherwise, it is in the form $S(x, y) = 0$, which is known as an implicit solution. For example, $y = 4e^{-x/2}$ is an explicit solution of $2y' + y = 0$. It turns out that a single formula $y = ke^{-x/2}$ involving a constant $k \neq 0$ generates all solutions of this ODE. Such formula is referred to as a general solution, and the constant is known as the parameter. When a specific value is assigned to the parameter, a particular solution is obtained.

Initial-Value Problem (IVP)

A first-order initial-value problem (IVP) appears in the form

$$y' = f(x, y) , \quad y(x_0) = y_0 , \tag{1.28}$$

where $y(x_0) = y_0$, is called the initial condition.

Example 1.8: IVP
Solve the initial-value problem

$$2y' + y = 0 , \quad y(2) = 3 .$$

Solution. As mentioned earlier, a general solution is $y = ke^{-x/2}$. Applying the initial condition, we obtain

$$y(2) = ke^{-1} = 3 \overset{\text{Solve for k}}{\Rightarrow} k = 3e .$$

Therefore, the particular solution is $y = 3e \cdot e^{-x/2} = 3e^{1-x/2}$.

Separable First-Order Ordinary Differential Equations

A first-order ODE is referred to as separable if it can be written as

$$f(y)y' = g(x)\,. \tag{1.29}$$

Using $y' = \mathrm{d}y/\mathrm{d}x$ in (1.29), we have

$$f(y)\frac{\mathrm{d}y}{\mathrm{d}x} = g(x) \Rightarrow f(y)\,\mathrm{d}y = g(x)\,\mathrm{d}x\,. \tag{1.30}$$

Integrating the two sides of (1.30) separately, yields

$$\int f(y)\,\mathrm{d}y = \int g(x)\,\mathrm{d}x + c\,, \quad c = \text{const.}$$

Example 1.9: Separable ODE
Solve the initial-value problem $\mathrm{e}^x y' = y^2$, $y(0) = 1$.

Solution. The ODE is separable and treated as

$$\begin{aligned}
&\mathrm{e}^x \frac{\mathrm{d}y}{\mathrm{d}x} = y^2 \\
\overset{\text{Provided that } y \neq 0}{\Rightarrow} \quad &\int \frac{1}{y^2}\,\mathrm{d}y = \int \frac{1}{\mathrm{e}^x}\,\mathrm{d}x \\
\Rightarrow \quad &-\frac{1}{y} = -\mathrm{e}^{-x} + c \qquad (c = \text{const.}) \\
\overset{\text{Solve for } y}{\Rightarrow} \quad &y(x) = \frac{1}{\mathrm{e}^{-x} - c}\,,
\end{aligned}$$

which is the general solution to the original differential equation. The specific value of c is determined via the given initial condition, as

$$\left.\begin{aligned} y(0) &= 1 \\ y(0) &= \tfrac{1}{1-c} \end{aligned}\right\} \begin{matrix} \text{Initial condition} \\ \\ \text{By gen. solution} \end{matrix} \Rightarrow \frac{1}{1-c} = 1 \Rightarrow c = 0\,.$$

Substitution into the general solution yields the particular solution $y(x) = \mathrm{e}^x$.

Linear First-Order Ordinary Differential Equations

A differential equation that can be expressed in the form

$$y' + g(x)y = f(x)\,, \tag{1.31}$$

where g and f are given functions of x, is called a linear first-order ordinary ODE. This of course agrees with what was discussed in (1.25) with slight changes in notation. If $f(x) = 0$ for every x in the interval under consideration, that is, if f is identically zero, denoted $f(x) \equiv 0$, then the ODE is called homogeneous, otherwise it is called nonhomogeneous.

Solution of Linear First-Order ODEs

The general solution of (1.31) can be expressed as [1.1, 4]

$$y(x) = \mathrm{e}^{-h(x)}\left[\int \mathrm{e}^{h(x)} f(x)\,\mathrm{d}x + c\right]\,,$$

$$\text{where} \quad h(x) = \int g(x)\,\mathrm{d}x\,. \tag{1.32}$$

Note that the constant of integration in the calculation of h is omitted because c accounts for all constants.

Example 1.10: Linear first-order ODE
Find the particular solution to the initial-value problem $2\dot{y} + y = 4\mathrm{e}^{2t}$, $y(0) = 1$.

Solution. Noting that t is now the independent variable, we first rewrite the ODE to agree with the form of (1.31), as

$$\dot{y} + \frac{1}{2}y = 2\mathrm{e}^{2t}$$

so that

$$g = \frac{1}{2}\,, \quad f = 2\mathrm{e}^{2t}\,.$$

With $h = \int g(t)\,\mathrm{d}t = \int \frac{1}{2}\,\mathrm{d}t = \frac{1}{2}t$, the general solution is given by (1.32),

$$\begin{aligned}
y(t) &= \mathrm{e}^{-t/2}\left[\int \mathrm{e}^{t/2}\cdot 2\mathrm{e}^{2t}\,\mathrm{d}t + c\right] \\
&= \mathrm{e}^{-t/2}\left[2\int \mathrm{e}^{5t/2}\,\mathrm{d}t + c\right] = \frac{4}{5}\mathrm{e}^{2t} + c\,\mathrm{e}^{-t/2}\,.
\end{aligned}$$

Applying the initial condition, we find $y(0) = \frac{4}{5} + c = 1 \Rightarrow c = \frac{1}{5}$. The particular solution is $y(t) = \frac{4}{5}\mathrm{e}^{2t} + \frac{1}{5}\mathrm{e}^{-t/2}$.

1.2.2 Numerical Solution of First-Order Ordinary Differential Equations

Recall that a first-order ODE can appear in an implicit form $F(x, y, y') = 0$ or an explicit form $y' = f(x, y)$. We will consider the latter, and assume that it is subject to a prescribed initial condition, that is,

$$y' = f(x, y)\,, \quad y(x_0) = y_0\,, \quad x_0 \le x \le x_N\,. \tag{1.33}$$

If finding a closed-form solution of (1.33) is difficult or impossible, we resort to a numerical solution. What

this means is that we find approximate values for the solution $y(x)$ at several points

$$x_1 = x_0 + h\ ,\ x_2 = x_0 + 2h \cdots x_n$$
$$= x_0 + nh\ ,\cdots,\ x_N = x_0 + Nh$$

known as mesh points, where h is called the step size. Note that the mesh points are equally spaced. Among many numerical methods to solve (1.33), the fourth-order Runge–Kutta method is most commonly used in practice. The difference equation for the fourth-order Runge–Kutta method (RK4) is derived as [1.5, 6]

$$y_{n+1} = y_n + \frac{1}{6}(q_1 + 2q_2 + 2q_3 + q_4)\ , \qquad (1.34)$$
$$n = 0,\ 1,\cdots,\ N-1\ ,$$

where

$$q_1 = h f(x_n,\ y_n)$$
$$q_2 = h f\left(x_n + \frac{h}{2},\ y_n + \frac{q_1}{2}\right)\ ,$$
$$q_3 = h f\left(x_n + \frac{h}{2},\ y_n + \frac{q_2}{2}\right)\ ,$$
$$q_4 = h f(x_n + h,\ y_n + q_3)\ .$$

Example 1.11: Fourth-order Runge–Kutta method
Apply RK4 with step size $h = 0.1$ to solve $y' + y = 2x^2$, $y(0) = 3$, $0 \le x \le 1$.

Solution. Knowing that $f(x_n,\ y_n) = -y_n + 2x_n^2$, the four function evaluations/step of the RK4 are

$$q_1 = h\left[-y_n + 2x_n^2\right]\ ,$$
$$q_2 = h\left[-\left(y_n + \frac{1}{2}q_1\right) + 2\left(x_n + \frac{1}{2}h\right)^2\right]\ ,$$
$$q_3 = h\left[-\left(y_n + \frac{1}{2}q_2\right) + 2\left(x_n + \frac{1}{2}h\right)^2\right]\ ,$$
$$q_4 = h\left[-(y_n + q_3) + 2(x_n + h)^2\right]\ .$$

Upon completion of each step, y_{n+1} is calculated by (1.34). So, we start with $n = 0$, corresponding to $x_0 = 0$ and $y_0 = 3$, and continue the process up to $n = 10$. Numerical results are generated as

$$y(0) = 3\ ,$$
$$y(0.1) = 2.7152\ ,$$
$$y(0.2) = 2.4613\ ,$$
$$y(0.3) = 2.2392\ ,\cdots,$$
$$y(0.9) = 1.6134\ ,$$
$$y(1) = 1.6321\ .$$

Further inspection reveals that RK4 produces the exact values (at least to five-decimal place accuracy) of the solution at the mesh points.

1.2.3 Second- and Higher-Order, Ordinary Differential Equations

The application of basic laws such as Newton's second law and Kirchhoff's voltage law (KVL) leads to mathematical models that are described by second-order ODEs [1.3]. Although it is quite possible that the system models contain nonlinear elements, in this section we will mainly focus on linear second-order differential equations. Nonlinear systems may be treated via numerical techniques such as the fourth-order Runge–Kutta method (Sect. 1.2), or via linearization [1.3]. In agreement with (1.25), a second-order ODE is said to be linear if it can be expressed in the form

$$y'' + g(x)y' + h(x)y = f(x)\ , \qquad (1.35)$$

where f, g, and h are given functions of x. Otherwise, it is called nonlinear.

Homogeneous Linear Second-Order ODEs

If y_1 and y_2 are two solutions of the homogeneous linear ODE

$$y'' + g(x)y' + h(x)y = 0 \qquad (1.36)$$

on some open interval, their linear combination $y = c_1 y_1 + c_2 y_2$ (c_1, c_2 constants) is also a solution on the same interval. This is known as the principle of superposition.

General Solution of Linear Second-Order ODEs – Linear Independence

A general solution of (1.36) is based on the idea of linear independence of functions, which involves what is known as the Wronskian. We first mention that a 2×2 determinant (Sect. 1.5.1) is evaluated as

$$\begin{vmatrix} p & q \\ r & s \end{vmatrix} = ps - qr\ .$$

If each of the functions $y_1(x)$ and $y_2(x)$ has at least a first derivative, then their Wronskian is denoted by $W(y_1,\ y_2)$ and is defined as the 2×2 determinant

$$W(y_1,\ y_2) = \begin{vmatrix} y_1 & y_2 \\ y_1' & y_2' \end{vmatrix} = y_1 y_2' - y_2 y_1'\ . \qquad (1.37)$$

If there exists a point $x^* \in (a,\ b)$ where $W \neq 0$, then y_1 and y_2 are linearly independent on the entire interval $(a,\ b)$.

Example 1.12: Independent solutions – the Wronskian
The functions $y_1 = e^{2x}$ and $y_2 = e^{-3x}$ are linearly independent for all x because their Wronskian is

$$W(y_1, y_2) = \begin{vmatrix} y_1 & y_2 \\ y_1' & y_2' \end{vmatrix} = \begin{vmatrix} e^{2x} & e^{-3x} \\ 2e^{2x} & -3e^{-3x} \end{vmatrix} = -5e^{-x} \neq 0$$

for all x.

If y_1 and y_2 are two linearly independent solutions of (1.36) on the interval (a, b), they form a basis of solutions for (1.36) on (a, b). A general solution of (1.36) on (a, b) is a linear combination of the basis elements, that is,

$$y = c_1 y_1 + c_2 y_2 \quad (c_1, c_2 \text{ constants}) . \tag{1.38}$$

Example 1.13: General solution, basis
It can be easily verified that $y_1 = e^{2x}$ and $y_2 = e^{-3x}$ are solutions of $y'' + y' - 6y = 0$ for all x. They are also linearly independent by Example 1.12. Consequently, $y_1 = e^{2x}$ and $y_2 = e^{-3x}$ form a basis of solutions for the ODE at hand, and a general solution for this ODE is $y = c_1 e^{2x} + c_2 e^{-3x}$ (c_1, c_2 constants).

Example 1.14: Unique solution of an IVP
Find the particular solution of $y'' + y' - 6y = 0$, $y(0) = -1$, $y'(0) = 8$.

Solution. By Example 1.13, a general solution is $y = c_1 e^{2x} + c_2 e^{-3x}$. Differentiating and applying the initial conditions, we have

$$\begin{aligned} y(0) &= c_1 + c_2 = -1 \\ y'(0) &= 2c_1 - 3c_2 = 8 \end{aligned} \quad \overset{\text{Solve the system}}{\Rightarrow} \quad \begin{aligned} c_1 &= 1 \\ c_2 &= -2 \end{aligned} .$$

Therefore, the unique solution of the IVP is obtained as $y = e^{2x} - 2e^{-3x}$.

Homogeneous Second-Order Differential Equations with Constant Coefficients

Consider a homogeneous linear second-order ODE with constant coefficients,

$$y'' + a_1 y' + a_2 y = 0 \quad (a_1, a_2 \text{ constants}) \tag{1.39}$$

and assume that its solution is in the form $y = e^{\lambda x}$, where λ, known as the *characteristic value*, is to be determined. Substitution into (1.39), yields

$$\lambda^2 e^{\lambda x} + a_1 \lambda e^{\lambda x} + a_2 e^{\lambda x} = 0$$
$$\Rightarrow \; e^{\lambda x}\left(\lambda^2 + a_1 \lambda + a_2\right) = 0 .$$

Since $e^{\lambda x} \neq 0$ for any finite values of x and λ, then

$$\lambda^2 + a_1 \lambda + a_2 = 0$$
$$\overset{\text{Solve the characteristic equation}}{\Rightarrow} \quad \begin{aligned} \lambda_1 &= \tfrac{1}{2}\left(-a_1 + \sqrt{a_1^2 - 4a_2}\right) \\ \lambda_2 &= \tfrac{1}{2}\left(-a_1 - \sqrt{a_1^2 - 4a_2}\right) \end{aligned} .$$

(1.40)

The solutions λ_1 and λ_2 of the characteristic equation are the characteristic values. The assumption was $y = e^{\lambda x}$, hence the solutions of (1.39) are $y_1 = e^{\lambda_1 x}$ and $y_2 = e^{\lambda_2 x}$. To find a general solution of (1.39), the two independent solutions must be identified. But this depends on the nature of the characteristic values λ_1 and λ_2, as discussed below.

Case 1: Two Distinct Real Roots $\left(a_1^2 - 4a_2 > 0, \lambda_1 \neq \lambda_2\right)$.
In this case, the solutions $y_1 = e^{\lambda_1 x}$ and $y_2 = e^{\lambda_2 x}$ are linearly independent, as may easily be verified. Thus, they form a basis of solution for (1.39). Therefore, a general solution is

$$y(x) = \underbrace{c_1 e^{\lambda_1 x} + c_2 e^{\lambda_2 x}}_{\text{General solution} - \lambda_1 \neq \lambda_2 \text{ , real}} . \tag{1.41}$$

Case 2: Double (Real) Root $\left(a_1^2 - 4a_2 = 0, \; \lambda_1 = \lambda_2 = -\tfrac{1}{2}a_1\right)$. It can be shown [1.1] that the two linearly independent solutions are $y_1 = e^{-a_1 x/2}$ and $y_2 = x e^{-a_1 x/2}$. Therefore,

$$\begin{aligned} y(x) &= c_1 e^{-\frac{1}{2}a_1 x} + c_2 x e^{-\frac{1}{2}a_1 x} \\ &= \underbrace{(c_1 + c_2 x) e^{-\frac{1}{2}a_1 x}}_{\text{General solution} - \lambda_1 = \lambda_2 \text{ , real}} . \end{aligned} \tag{1.42}$$

Case 3: Complex Conjugate Pair $\left(a_1^2 - 4a_2 < 0, \lambda_1 = \bar{\lambda}_2\right)$.
The characteristic values are given as $\lambda_{1,2} = \frac{1}{2}(-a_1 \pm \sqrt{a_1^2 - 4a_2})$. Since $a_1^2 - 4a_2 < 0$, we write

$$\begin{aligned} \lambda_{1,2} &= \frac{1}{2}\left[-a_1 \pm \sqrt{-\left(4a_2 - a_1^2\right)}\right] \\ &= \frac{1}{2}\left[-a_1 \pm \sqrt{-1}\sqrt{4a_2 - a_1^2}\right] \\ &= \frac{1}{2}\left[-a_1 \pm i\sqrt{4a_2 - a_1^2}\right] = -\sigma \pm i\omega , \end{aligned}$$

$(i = \sqrt{-1})$

where

$$\sigma = \frac{1}{2}a_1 \,,$$
$$\omega = \frac{1}{2}\sqrt{4a_2 - a_1^2} \,. \tag{1.43}$$

The two independent solutions are $y_1 = \mathrm{e}^{-\sigma x}\cos\omega x$ and $y_2 = \mathrm{e}^{-\sigma x}\sin\omega x$, and a general solution of (1.39) is obtained as

$$y(x) = \underbrace{\mathrm{e}^{-\sigma x}(c_1\cos\omega x + c_2\sin\omega x)}_{\text{General solution — } \lambda_1 = \bar{\lambda}_2 \text{ , complex conjugates}} \,. \tag{1.44}$$

Example 1.15: Case (3)
Solve $y'' + 2y' + 2y = 0$, $y(0) = 1$, $y'(0) = 0$.

Solution. We first find the characteristic equation and the corresponding characteristic values, as

$$\lambda^2 + 2\lambda + 2 = 0$$
$$\Rightarrow \lambda_{1,2} = -1 \pm \mathrm{i} \,.$$

Complex conjugate pair, Case (3)

By (1.43), we identify $\sigma = 1$ and $\omega = 1$, so that the general solution by (1.44) is

$$y(x) = \mathrm{e}^{-x}(c_1\cos x + c_2\sin x) \,.$$

Next, we differentiate this to obtain

$$y'(x) = -\mathrm{e}^{-x}(c_1\cos x + c_2\sin x) + \mathrm{e}^{-x}(-c_1\sin x + c_2\cos x)$$

Finally, by the initial conditions,

$$\begin{aligned} y(0) &= c_1 = 1 \\ y'(0) &= -c_1 + c_2 = 0 \end{aligned} \quad \Rightarrow \quad \begin{aligned} c_1 &= 1 \\ c_2 &= 1 \end{aligned}$$

and the solution is $y(x) = \mathrm{e}^{-x}(\cos x + \sin x)$.

Boundary-Value Problems (BVP). In certain applications involving second-order differential equations, a pair of information is provided at the boundary points of an open interval (a, b) on which the ODE is to be solved. This pair is referred to as the boundary conditions, and the problem

$$y'' + a_1 y' + a_2 y = 0 \,,$$
$$\underbrace{y(a) = A \,,\; y(b) = B}_{\text{Boundary conditions}} \tag{1.45}$$

is called a boundary-value problem (BVP).

Nonhomogeneous Linear Second-Order ODEs

Nonhomogeneous second-order ODEs appear in the form

$$y'' + g(x)y' + h(x)y = f(x) \,, \qquad f(x) \not\equiv 0 \,. \tag{1.46}$$

A general solution for this equation is then obtained as

$$y(x) = \underbrace{y_\mathrm{h}(x)}_{\text{Homogeneous solution}} + \underbrace{y_\mathrm{p}(x)}_{\text{Particular solution}} \,. \tag{1.47}$$

Homogeneous Solution $y_\mathrm{h}(x)$. $y_\mathrm{h}(x)$ is a general solution of the homogeneous equation (1.36), and as previously discussed, it is given by

$$y_\mathrm{h} = c_1 y_1 + c_2 y_2 \,, \quad (c_1, c_2 \text{ constants})$$

where y_1 and y_2 are linearly independent and form a basis of solutions for (1.36). Note that the homogeneous solution involves two arbitrary constants.

Particular Solution $y_\mathrm{p}(x)$. $y_\mathrm{p}(x)$ is a particular solution of (1.46), and does not involve any arbitrary constants. The nature of $y_\mathrm{p}(x)$ depends on the nature of $f(x)$, as well as its relation to the independent solutions y_1 and y_2 of the homogeneous equation.

Method of Undetermined Coefficients

When (1.46) happens to have constant coefficients and the function $f(x)$ is of a special type – polynomial, exponential, sine and/or cosine or a combination of them – then the particular solution can be obtained by the method of undetermined coefficients as follows. Consider

$$y'' + a_1 y' + a_2 y = f(x) \quad (a_1, a_2 \text{ constants}) \,. \tag{1.48}$$

Since the coefficients are constants, the homogeneous solution y_h is found as before. So all we need to do is to find the particular solution y_p. We will make a proper selection for y_p based on the nature of $f(x)$ and with the

Table 1.1 Selection of particular solution – the method of undetermined coefficients

Term in $f(x)$	Proper choice of y_p
$a_n x^n + \ldots + a_1 x + a_0$	$K_n x^n + \ldots + K_1 x + K_0$
$A\,\mathrm{e}^{ax}$	$K\,\mathrm{e}^{ax}$
$A\sin\omega x$ or $A\cos\omega x$	$K_1\cos\omega x + K_2\sin\omega x$
$A\,\mathrm{e}^{\sigma x}\sin\omega x$ or $A\,\mathrm{e}^{\sigma x}\cos\omega x$	$\mathrm{e}^{\sigma x}(K_1\cos\omega x + K_2\sin\omega x)$

aid of Table 1.1. This choice involves unknown coefficients, which will be determined by substituting y_p and its derivatives into (1.48). The details, as well as special cases that may occur, are given below.

Procedure.

Step 1: Homogeneous Solution $y_h(x)$. Solve the homogeneous equation $y'' + a_1 y' + a_2 y = 0$ to find the two independent solutions y_1 and y_2, and the general solution $y_h(x) = c_1 y_1(x) + c_2 y_2(x)$.

Step 2: Particular Solution $y_p(x)$. For each term in $f(x)$ choose a proper y_p as suggested by Table 1.1. For instance, if $f(x) = x + 2e^x$ then pick $y_p = K_1 x + K_2 + K e^x$. Note that, if instead of x we had $3x - 2$, for example, the choice of y_p would still be the same because they both represent first-degree polynomials. We then substitute our choice of y_p, along with its derivatives, into the original ODE to find the undetermined coefficients.

Special cases.

I. Suppose a term in our choice of y_p coincides with a solution (y_1 or y_2) of the homogeneous equation, and that this solution is associated with a *simple* (i. e., nonrepeated) characteristic value. Then, make the modification by multiplying y_p by x.

II. If a term in the choice of y_p coincides with a solution of the homogeneous equation, and that this solution is associated with a *repeated* characteristic value, modify by multiplying y_p by x^2.

Example 1.16: Special case II

Solve

$$y'' + 2y' + y = x + 1 + 3e^{-x}, \quad y(0) = 1, \quad y'(0) = 0.$$

Step 1: Homogeneous Solution. The characteristic equation $(\lambda + 1)^2 = 0$ yields a *double root* $\lambda = -1$. This means $y_1 = e^{-x}$ and $y_2 = x e^{-x}$, so that the homogeneous solution is $y_h(x) = (c_1 + c_2 x)e^{-x}$.

Step 2: Particular Solution. The right-hand side of the ODE consists of two functions,

$$\underbrace{x + 1}_{\text{First-degree polynomial}} \quad \text{and} \quad e^{-x}.$$

The first term, $x + 1$, does not coincide with either y_1 or y_2, so the proper choice by Table 1.1 is $K_1 x + K_0$. The second term involves e^{-x}, which happens to be a homogeneous solution associated with a double root. Therefore, by special case II the modified choice is $K x^2 e^{-x}$. Consequently, the particular solution is in the form

$$y_p(x) = \underbrace{K_1 x + K_0}_{\text{First term}} + \underbrace{K x^2 e^{-x}}_{\text{Second term}}.$$

Substitution of y_p and its derivatives into the nonhomogeneous ODE, and collecting terms, results in

$$2K e^{-x} + K_1 x + K_0 + 2K_1 = x + 1 + 3e^{-x}.$$

Equating the coefficients of like terms, we have

$$\begin{aligned} 2K &= 3 && K = \tfrac{3}{2} \\ K_1 &= 1 \quad \Rightarrow && K_1 = 1 \\ K_0 + 2K_1 &= 1 && K_0 = -1 \end{aligned}$$

$$\Rightarrow y_p(x) = x - 1 + \frac{3}{2} x^2 e^{-x}.$$

Step 3: General Solution. The general solution is then found as

$$y(x) = (c_1 + c_2 x)e^{-x} + x - 1 + \frac{3}{2} x^2 e^{-x}.$$

Step 4: Initial Conditions. Applying the initial conditions, we obtain $c_1 = 2$ and $c_2 = 1$. Finally, the solution to the IVP is

$$y(x) = (2 + x)e^{-x} + x - 1 + \frac{3}{2} x^2 e^{-x}.$$

Higher-Order Ordinary Differential Equations

Many of the techniques for the treatment of differential equations of order three or higher are merely extensions of those applied to second-order equations. Here we will only discuss nth-order, linear nonhomogeneous ODEs with constant coefficients, that is,

$$y^{(n)} + a_{n-1} y^{(n-1)} + \cdots + a_1 y' + a_0 y = f(x), \tag{1.49}$$

where $a_0, a_1, \cdots, a_{n-1}$ are constants. As in the case of second-order ODEs, a general solution consists of the homogeneous solution and the particular solution. For cases when $f(x)$ is of a special type, the particular solution is obtained via the method of undetermined coefficients.

Method of Undetermined Coefficients. The idea introduced for second-order ODEs is now extended to find y_p for (1.49). As before, a proper choice of y_p is made assuming that $f(x)$ consists of terms that are listed in Table 1.1. If none of the terms in $f(x)$ happens to be

an independent homogeneous solution, then no modification is necessary. Otherwise, the following special cases need be taken into account.

Special Cases.

1. If a term in our choice of y_p coincides with a homogeneous solution, which corresponds to a *simple* (nonrepeated) characteristic value, then we make the modification by multiplying y_p by x.
2. If a term in y_p coincides with a solution of the homogeneous equation, and this solution is associated with a characteristic value of *multiplicity* m, we modify by multiplying y_p by x^m.

Example 1.17: Special case II
Solve

$$y''' - 4y'' = 1 + 12x\,,$$
$$y(0) = 0\,, \quad y'(0) = 4\,, \quad y''(0) = 15\,.$$

Solution.
Step 1: Homogeneous Solution. Characteristic equation:

$$\lambda^3 - 4\lambda^2 = \lambda^2(\lambda - 4) = 0 \Rightarrow \lambda = 0, 0, 4\,.$$

Therefore $y_h = c_1 + c_2 x + c_3 e^{4x}$.

Step 2: Particular Solution. Noting that $f(x) = 1 + 12x$ is a first-degree polynomial, we pick $y_p = K_1 x + K_0$. But x happens to be a homogeneous solution associated with a double root ($\lambda = 0$). Hence, the modification is $y_p = (K_1 x + K_0)x^2$. Substituting this and its derivatives into the original ODE, and simplifying, we arrive at

$$(6K_1 - 8K_0) - 24K_1 x = 1 + 12x$$
$$\Rightarrow \begin{array}{l} 6K_1 - 8K_0 = 1 \\ -24K_1 = 12 \end{array} \Rightarrow \begin{array}{l} K_1 = -\frac{1}{2} \\ K_0 = -\frac{1}{2} \end{array}$$
$$\Rightarrow y_p = -\frac{1}{2}(x+1)x^2$$

Step 3: General Solution. Combination of y_h and y_p gives a general solution $y = c_1 + c_2 x + c_3 e^{4x} - \frac{1}{2}(x+1)x^2$.

Step 4: Initial Conditions. Applying the initial conditions to the general solution and its derivatives, we obtain

$$\begin{array}{ll} y(0) = c_1 + c_3 = 0 & c_1 = -1 \\ y'(0) = c_2 + 4c_3 = 4 & \Rightarrow\ c_2 = 0 \\ y''(0) = 16c_3 - 1 = 15 & c_3 = 1 \end{array}$$
$$\Rightarrow y(x) = -1 + e^{4x} - \frac{1}{2}x^3 - \frac{1}{2}x^2\,.$$

1.3 Laplace Transformation

In Sect. 1.2 we mainly learned to solve linear time-invariant (LTI) ODEs without ever leaving the time domain. In this section we introduce a systematic approach to solve such ODEs in a more-expedient manner. The primary advantage gained here is that the arbitrary constants in the general solution need not be found separately. The idea is simple: in order to solve an ODE and corresponding initial-value problem (IVP) or boundary-value problem (BVP), transform the problem to the so-called s domain, in which the transformed problem is an algebraic one. This algebraic problem is then treated properly, and the data is ultimately transformed back to time domain to find the solution of the original problem. The transform function is a function of a complex variable, denoted by s. If a function $f(t)$ is defined for all $t \geq 0$, then its Laplace transform is defined by

$$F(s) \overset{\text{Notation}}{=} \mathbf{L}[f(t)]$$
$$\overset{\text{Definition}}{=} \int_0^\infty e^{-st} f(t)\,dt \qquad (1.50)$$

provided that the integral exists. The complex variable s is the Laplace variable, and $\mathbf{L}$ is the Laplace transform operator. It is common practice to denote a time-dependent function by a lower-case letter, say, $f(t)$, and its Laplace transform by the same letter in upper case, $F(s)$.

1.3.1 Inverse Laplace Transform

Suppose we are seeking the solution $x(t)$ of an ODE. The ODE is first transformed into the s domain by means of the operator $\mathbf{L}$. In this domain, the transformed version of the ODE is an algebraic equation involving the transform function $X(s)$ of $x(t)$. This equation is then manipulated to find $X(s)$, which in turn will be transformed back into time domain to determine $x(t)$. This is done through the inverse Laplace transformation, as in Fig. 1.9.

$$x(t) = \mathbf{L}^{-1}[X(s)]$$

Consistent with (1.50), we have

$$f(t) = \mathbf{L}^{-1}[F(s)] \,. \tag{1.51}$$

Example 1.18: Laplace transform
Given $g(t) = 1$ for $t \geq 0$, find $\mathbf{L}[g(t)]$.

Solution

Following the definition given by (1.50), we have

$$\mathbf{L}[g(t)] = \mathbf{L}(1) = \int_0^\infty \mathrm{e}^{-st}\,\mathrm{d}t = \left.\frac{\mathrm{e}^{-st}}{-s}\right|_{t=0}^{\infty} = \frac{1}{s}$$

for $s > 0$.

Example 1.19: Laplace transform
Suppose $h(t) = \mathrm{e}^{-at}$ (a = const) for $t \geq 0$. Determine $H(s)$.

Solution

By definition,

$$H(s) = \mathbf{L}[\mathrm{e}^{-at}] = \int_0^\infty \mathrm{e}^{-st}\,\mathrm{e}^{-at}\,\mathrm{d}t = \int_0^\infty \mathrm{e}^{-(s+a)t}\,\mathrm{d}t = \left.\frac{\mathrm{e}^{-(s+a)t}}{-(s+a)}\right|_{t=0}^{\infty} = \frac{1}{s+a}$$

for $s + a > 0$.

Linearity of Laplace and Inverse Laplace Transforms

The Laplace transform operator $\mathbf{L}$ is linear, that is, if the Laplace transforms of functions $f_1(t)$ and $f_2(t)$ exist, and a_1 and a_2 are constant scalars, then

$$\begin{aligned}
&\mathbf{L}[a_1 f_1(t) + a_2 f_2(t)] \\
&= \int_0^\infty \mathrm{e}^{-st}[a_1 f_1(t) + a_2 f_2(t)]\,\mathrm{d}t \\
&= a_1 \int_0^\infty \mathrm{e}^{-st} f_1(t)\,\mathrm{d}t + a_2 \int_0^\infty \mathrm{e}^{-st} f_2(t)\,\mathrm{d}t \\
&= a_1 \mathbf{L}[f_1(t)] + a_2 \mathbf{L}[f_2(t)] \\
&= a_1 F_1(s) + a_2 F_2(s) \,.
\end{aligned} \tag{1.52}$$

To establish the linearity of $\mathbf{L}^{-1}$, take the inverse Laplace transforms of the expressions on the far left and far right of (1.52) to obtain

$$a_1 f_1(t) + a_2 f_2(t) = \mathbf{L}^{-1}[a_1 F_1(s) + a_2 F_2(s)] \,.$$

Noting that $f_1(t) = \mathbf{L}^{-1}[F_1(s)]$ and $f_2(t) = \mathbf{L}^{-1}[F_2(s)]$, the result follows.

Example 1.20: Linearity of $\mathbf{L}$
Find $\mathbf{L}[2 - 3\mathrm{e}^{4t}]$.

Solution. Using the linearity of $\mathbf{L}$, we write $\mathbf{L}[2 - 3\mathrm{e}^{4t}] = 2\mathbf{L}[1] - 3L[\mathrm{e}^{4t}]$. But, by Example 1.18 we have $\mathbf{L}[1] = 1/s$. And by Example 1.18 (with $a = -4$) we have $\mathbf{L}[\mathrm{e}^{4t}] = 1/(s-4)$. Thus $\mathbf{L}[2 - 3\mathrm{e}^{4t}] = 2/s - 3/(s-4) = (-(s+8))/(s(s-4))$.

Table of Laplace Transform Pairs. Laplace transforms of several functions are listed in Table 1.2 at the end of this section. We will refer to this frequently. For a better understanding of the concepts, however, we try to derive the most fundamental results on our own.

Theorem 1.1: Shift on the s-axis. Suppose that $F(s) = \mathbf{L}[f(t)]$ and that a is a constant. Then,

$$\mathbf{L}[\mathrm{e}^{-at} f(t)] = F(s+a) \,. \tag{1.53}$$

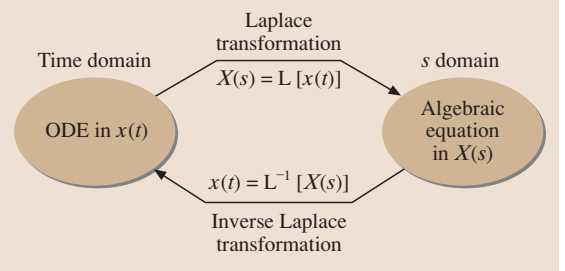


Fig. 1.9 Operations involved in the Laplace transformation method

Table 1.2 Laplace transform pairs

No.	$f(t)$	$F(s)$
1	Unit impulse $\delta(t)$	1
2	1, unit step $u_s(t)$	$1/s$
3	t, unit ramp $u_r(t)$	$1/s^2$
4	$\delta(t-a)$	e^{-as}
5	$u(t-a)$	e^{-as}/s
6	t^{n-1}, $n=1,2,\ldots$	$(n-1)!/s^n$
7	t^{a-1}, $a>0$	$\Gamma(a)/s^a$
8	e^{-at}	$\frac{1}{s+a}$
9	te^{-at}	$\frac{1}{(s+a)^2}$
10	$t^n e^{-at}$, $n=1,2,\ldots$	$\frac{n!}{(s+a)^{n+1}}$
11	$\frac{1}{b-a}(e^{-at}-e^{-bt})$, $a\neq b$	$\frac{1}{(s+a)(s+b)}$
12	$\frac{1}{a-b}(ae^{-at}-be^{-bt})$, $a\neq b$	$\frac{s}{(s+a)(s+b)}$
13	$\frac{1}{ab}\left[1+\frac{1}{a-b}(be^{-at}-ae^{-bt})\right]$	$\frac{1}{s(s+a)(s+b)}$
14	$\frac{1}{a^2}(-1+at+e^{-at})$	$\frac{1}{s^2(s+a)}$
15	$\frac{1}{a^2}(1-e^{-at}-ate^{-at})$	$\frac{1}{s(s+a)^2}$
16	$\sin\omega t$	$\frac{\omega}{s^2+\omega^2}$
17	$\cos\omega t$	$\frac{s}{s^2+\omega^2}$
18	$e^{-\sigma t}\sin\omega t$	$\frac{\omega}{(s+\sigma)^2+\omega^2}$
19	$e^{-\sigma t}\cos\omega t$	$\frac{\omega^2}{s(s^2+\omega^2)}$
20	$1-\cos\omega t$	$\frac{s+\sigma}{(s+\sigma)^2+\omega^2}$
21	$\omega t-\sin\omega t$	$\frac{\omega^3}{s^2(s^2+\omega^2)}$
22	$t\cos\omega t$	$\frac{s^2-\omega^2}{(s^2+\omega^2)^2}$
23	$\frac{1}{2\omega}t\sin\omega t$	$\frac{s}{(s^2+\omega^2)^2}$
24	$\frac{1}{2\omega^3}(\sin\omega t-\omega t\cos\omega t)$	$\frac{1}{(s^2+\omega^2)^2}$
25	$\frac{1}{2\omega}(\sin\omega t+\omega t\cos\omega t)$	$\frac{s^2}{(s^2+\omega^2)^2}$
26	$\frac{1}{\omega_2^2-\omega_1^2}\left[\frac{1}{\omega_2}\sin\omega_2 t-\frac{1}{\omega_1}\sin\omega_1 t\right]$, $\omega_1^2\neq\omega_2^2$	$\frac{1}{(s^2+\omega_1^2)(s^2+\omega_2^2)}$
27	$\frac{1}{\omega_2^2-\omega_1^2}(\cos\omega_1 t-\cos\omega_2 t)$, $\omega_1^2\neq\omega_2^2$	$\frac{s}{(s^2+\omega_1^2)(s^2+\omega_2^2)}$
28	$\sinh at$	$\frac{a}{s^2-a^2}$
29	$\cosh at$	$\frac{s}{s^2-a^2}$
30	$\frac{1}{a^2-b^2}\left[\frac{1}{a}\sinh at-\frac{1}{b}\sinh bt\right]$, $a\neq b$	$\frac{1}{(s^2-a^2)(s^2-b^2)}$
31	$\frac{1}{a^2-b^2}[\cosh at-\cosh bt]$, $a\neq b$	$\frac{s}{(s^2-a^2)(s^2-b^2)}$

Table 1.2 Laplace transform pairs, continued

No.	$f(t)$	$F(s)$
32	$\frac{1}{3a^2}\left[e^{-at}+2e^{\frac{1}{2}at}\sin\left(\frac{\sqrt{3}}{2}at-\frac{\pi}{6}\right)\right]$	$\frac{1}{s^3-a^3}$
33	$\frac{1}{3a}\left[-e^{-at}+2e^{\frac{1}{2}at}\sin\left(\frac{\sqrt{3}}{2}at+\frac{\pi}{6}\right)\right]$	$\frac{s}{s^3-a^3}$
34	$\frac{1}{3a^2}\left[e^{at}-2e^{-\frac{1}{2}at}\sin\left(\frac{\sqrt{3}}{2}at+\frac{\pi}{6}\right)\right]$	$\frac{1}{s^3-a^3}$
35	$\frac{1}{3a}\left[e^{-at}+2e^{-\frac{1}{2}at}\sin\left(\frac{\sqrt{3}}{2}at-\frac{\pi}{6}\right)\right]$	$\frac{s}{s^3-a^3}$
36	$\frac{1}{4a^3}(\cosh at\sin at-\sinh at\cos at)$	$\frac{1}{s^4+4a^4}$
37	$\frac{1}{2a^2}\sinh at\sin at$	$\frac{s}{s^4+4a^4}$
38	$\frac{1}{2a^3}(\sinh at-\sin at)$	$\frac{1}{s^4-a^4}$
39	$\frac{1}{2a^2}(\cosh at-\cos at)$	$\frac{s}{s^4-a^4}$

See Fig. 1.10. Alternatively, in terms of the inverse Laplace transform,

$$\mathbf{L}^{-1}[F(s+a)] = e^{-at} f(t)\,. \tag{1.54}$$

Example 1.21: Shift on the s-axis
Find $\mathbf{L}[e^{3t}\cos t]$.

Solution. Let $f(t)=\cos t$ so that $F(s)=s/(s^2+1)$; see Table 1.2. Then, by (1.53) with $a=-3$,

$$\mathbf{L}[e^{3t}\cos t] \overset{\substack{f(t)=\cos t\\ a=-3}}{=} F(s-3) = \frac{s-3}{(s-3)^2+1}$$

Differentiation and Integration of Laplace Transforms

We now turn our attention to two specific types of situations: (1) $\mathbf{L}[tf(t)]$, (2) $\mathbf{L}[f(t)/t]$. In both cases,

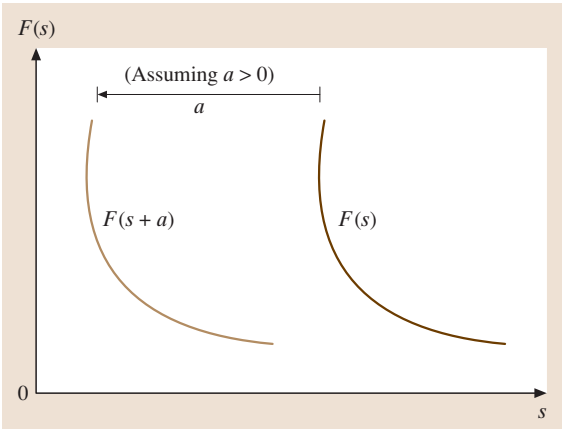


Fig. 1.10 Shift on the s-axis (Theorem 1.1)

we assume that $f(t)$ is such that $F(s) = \mathbf{L}[f(t)]$ is either known directly from Table 1.2 or can be determined by other means. Either way, once $F(s)$ is available, the two transforms labeled (1) and (2) will be obtained in terms of the derivative and integral of $F(s)$, respectively. Before presenting two key results pertaining to these situations we make the following definition. If a transform function is in the form $F(s) = N(s)/D(s)$, then each value of s for which $D(s) = 0$ is called a pole of $F(s)$. A pole with a *multiplicity* (number of occurrences) of one is known as a simple pole.

Theorem 1.2: Differentiation of Laplace Transforms. If $\mathbf{L}[f(t)] = F(s)$ exists, then at any point except at the poles of $F(s)$, we have

$$\mathbf{L}[tf(t)] = -\frac{\mathrm{d}}{\mathrm{d}s}F(s) = -F'(s) \tag{1.55}$$

or alternatively,

$$tf(t) = -\mathbf{L}^{-1}[F'(s)]\,. \tag{1.56}$$

The general form of (1.55) for $n = 1, 2, 3, \cdots$ is given by

$$\mathbf{L}[t^n f(t)] = (-1)^n \frac{\mathrm{d}^n}{\mathrm{d}s^n}F(s) = (-1)^n F^{(n)}(s)\,. \tag{1.57}$$

Example 1.22: Differentiation of $F(s)$
Find $\mathbf{L}[t \sin 3t]$.

Solution. Comparing with the left side of (1.55), we have $f(t) = \sin 3t$ so that $F(s) = 3/(s^2+9)$. Therefore,

$$\begin{aligned}\mathbf{L}[t \sin 3t] &= -\frac{\mathrm{d}}{\mathrm{d}s}\left(\frac{3}{s^2+9}\right)\\ &= \frac{6s}{(s^2+9)^2}\end{aligned}$$

Theorem 1.3: Integration of Laplace transforms. If $\mathbf{L}[f(t)/t]$ exists, and the order of integration can be interchanged, then

$$\mathbf{L}\left[\frac{f(t)}{t}\right] = \int_s^\infty F(\sigma)\,\mathrm{d}\sigma\,. \tag{1.58}$$

Alternatively,

$$f(t) = t\mathbf{L}^{-1}\left[\int_s^\infty F(\sigma)\,\mathrm{d}\sigma\right]\,. \tag{1.59}$$

Example 1.23: Theorem 1.3
Show that

$$\mathbf{L}\left[\frac{\sin \omega t}{t}\right] = \cot^{-1}\frac{s}{\omega}\,.$$

Solution. Comparing with (1.58), $f(t) = \sin \omega t$ so that $F(s) = \omega/(s^2+\omega^2)$. Subsequently,

$$\begin{aligned}\mathbf{L}\left[\frac{\sin \omega t}{t}\right] &= \int_s^\infty \frac{\omega}{\sigma^2+\omega^2}\,\mathrm{d}\sigma\\ &= \int_s^\infty \frac{1}{1+(\sigma/\omega)^2}\frac{d\sigma}{\omega}\\ &= \left[\tan^{-1}\frac{\sigma}{\omega}\right]_{\sigma=s}^\infty\\ &= \frac{\pi}{2} - \tan^{-1}\frac{s}{\omega} = \cot^{-1}\frac{s}{\omega}\,.\end{aligned}$$

1.3.2 Special Functions

Much can be learned about the characteristics of a system based on its response to specific external disturbances. To perform the response analysis, these disturbances must first be mathematically modeled, which is where special functions play an important role. In this section we will introduce the step, ramp, pulse, and impulse functions, as well as their Laplace transforms.

Unit Step $u(t)$

The unit-step function (Fig. 1.11) is analytically defined as

$$u(t) = \begin{cases} 1 & \text{if} \quad t > 0\\ 0 & \text{if} \quad t < 0\\ \text{undefined (finite)} & \text{if} \quad t = 0\end{cases}\,. \tag{1.60}$$

This may be physically realized as a constant signal (of magnitude 1) suddenly applied to the system at time $t = 0$. By the definition of the Laplace transform, we find

$$\begin{aligned}\mathbf{L}[u(t)] \overset{\text{Notation}}{=} U(s) &= \int_0^\infty \mathrm{e}^{-st}u(t)\,\mathrm{d}t\\ &= \int_0^\infty \mathrm{e}^{-st}\,\mathrm{d}t = \frac{1}{s}\,.\end{aligned} \tag{1.61}$$

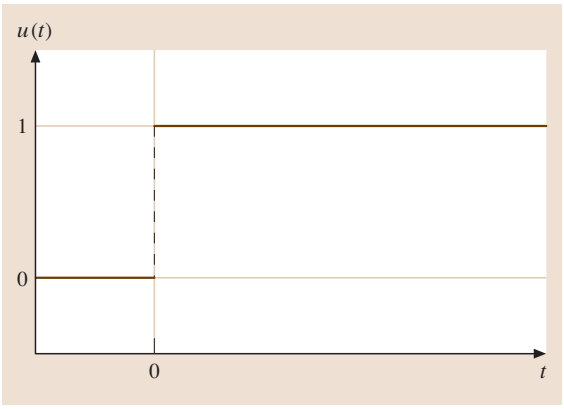


Fig. 1.11 The unit step function $u(t)$

When the magnitude is some $A \neq 1$, we refer to the signal as a step function, denoted by $Au(t)$. In this case,

$$\mathbf{L}[Au(t)] = \int_0^\infty \mathrm{e}^{-st} A \, \mathrm{d}t = \frac{A}{s} \, .$$

When the unit step function occurs at some time $a \neq 0$ (Fig. 1.12), it is denoted by $u(t-a)$, and

$$u(t-a) = \begin{cases} 1 & \text{if} \quad t > a \\ 0 & \text{if} \quad t < a \\ \text{undefined (finite)} & \text{if} \quad t = a \end{cases} \, . \tag{1.62}$$

As before, if the magnitude happens to be $A \neq 1$, the notation is modified to $Au(t-a)$. To find the Laplace transform of $u(t-a)$, we first need to discuss the shift on the t-axis, see Theorem 1.4 below.

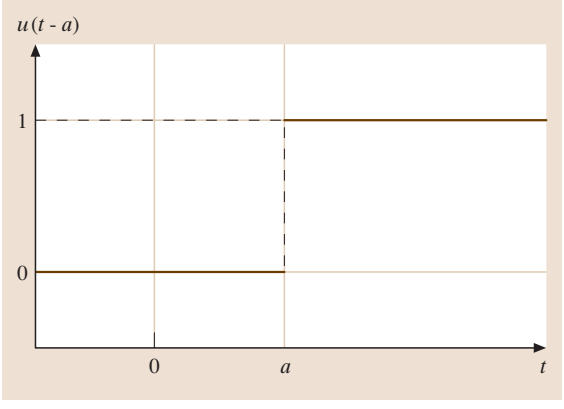


Fig. 1.12 The unit -step function occurring at $t = a$

Theorem 1.4: Shift on the t-axis. Given that $F(s) = \mathbf{L}[f(t)]$ exists, then

$$\mathbf{L}[f(t-a)u(t-a)] = \mathrm{e}^{-as} F(s) \, , \tag{1.63}$$

or, alternatively,

$$\mathbf{L}^{-1}[\mathrm{e}^{-as} F(s)] = f(t-a)u(t-a) \, . \tag{1.64}$$

Finding $\mathbf{L}[u(t-a)]$ via Theorem 1.4. We now have the tools to determine $\mathbf{L}[u(t-a)]$. In particular, comparing $\mathbf{L}[u(t-a)]$ with the left-hand side of (1.63), we deduce that $f(t-a) = 1$. Which implies that $f(t) = 1$, hence $F(s) = 1/s$. As a result,

$$\mathbf{L}[u(t-a)] = \frac{\mathrm{e}^{-as}}{s} \, . \tag{1.65}$$

Unit Ramp $u_\mathrm{r}(t)$. The unit ramp function (Fig. 1.13) is analytically defined as

$$u_\mathrm{r}(t) = \begin{cases} t & \text{if} \quad t \geq 0 \\ 0 & \text{if} \quad t < 0 \, . \end{cases}$$

Physically, this models a signal that changes linearly with a unit rate. By (1.50),

$$\begin{aligned} \mathbf{L}[u_\mathrm{r}(t)] \overset{\text{Notation}}{=} U_\mathrm{r}(s) &= \int_0^\infty t \, \mathrm{e}^{-st} \, \mathrm{d}t \\ &= \left(t \frac{\mathrm{e}^{-st}}{-s} \right)_{t=0}^{\infty} - \int_0^\infty \frac{\mathrm{e}^{-st}}{-s} \, \mathrm{d}t \\ &= \left[\frac{\mathrm{e}^{-st}}{-s^2} \right]_{t=0}^{\infty} = \frac{1}{s^2} \, . \end{aligned} \tag{1.66}$$

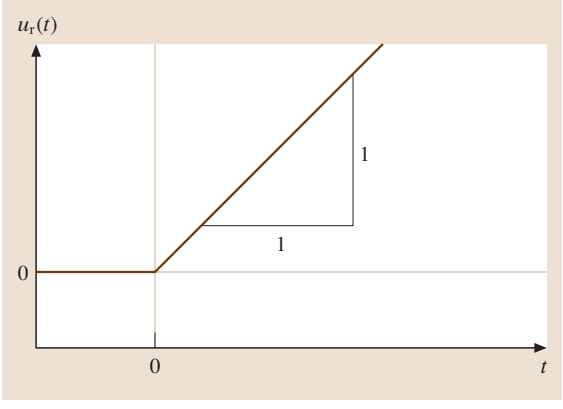


Fig. 1.13 The unit ramp function $u_\mathrm{r}(t)$

Note that $u_r(t) = tu(t)$. When the rate is $A \neq 1$, the signal is called a ramp function, denoted by $Au_r(t)$. In that case,

$$\mathbf{L}[Au_r(t)] = \frac{A}{s^2} .$$

Unit Pulse $u_p(t)$. The unit pulse function (Fig. 1.14) is defined as

$$u_p(t) = \begin{cases} 1/t_1 & \text{if} \quad 0 < t < t_1 \\ 0 & \text{if} \quad t < 0 \quad \text{and} \quad t > t_1 . \end{cases}$$

The word 'unit' signifies that the signal occupies an area of unity. Its Laplace transform is derived as

$$\mathbf{L}[u_p(t)] \overset{\text{Notation}}{=} U_p(s) = \int_0^{t_1} \frac{1}{t_1} e^{-st} \, dt = \frac{1 - e^{-st_1}}{st_1} . \qquad (1.67)$$

If the area is $A \neq 1$, the signal is called a pulse, written $Au_p(t)$, and

$$\mathbf{L}[Au_p(t)] = \frac{A(1 - e^{-st_1})}{st_1} .$$

Unit Impulse (Dirac Delta) $\delta(t)$. Consider the unit pulse of Fig. 1.14 and let $t_1 \to 0$; Fig. 1.15. In this limit, the rectangular-shaped signal occupies a region with an infinitesimally small width and a large height (Fig. 1.16). The area, however, remains unity throughout the process. This limiting signal is known as the unit impulse (or Dirac delta), denoted by $\delta(t)$. If the area is $A \neq 1$, it is an impulse, denoted by $A\delta(t)$. If an external disturbance (such as an applied force or voltage) is a pulse with very large magnitude and applied for a very short period of time, then it can be approximated as an impulse. Since $\delta(t)$ is the limit of $u_p(t)$ as $t_1 \to 0$, we

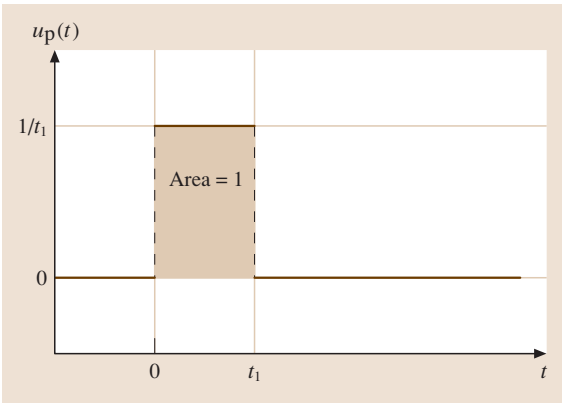


Fig. 1.14 The unit pulse $u_p(t)$

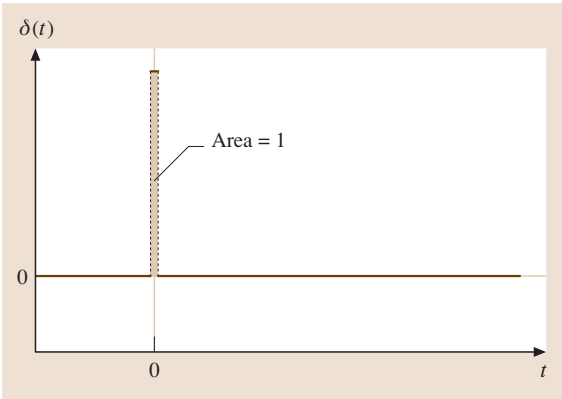


Fig. 1.16 The unit impulse $\delta(t)$

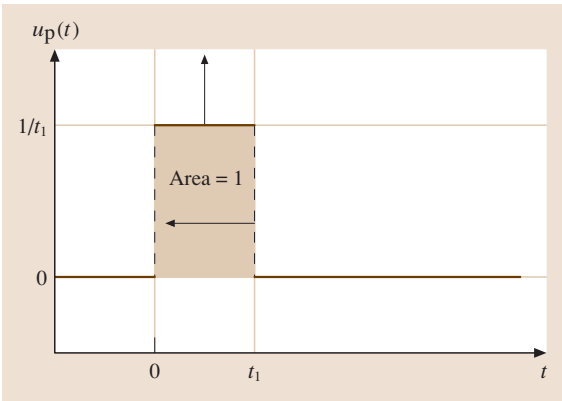


Fig. 1.15 The unit pulse as $t_1 \to 0$

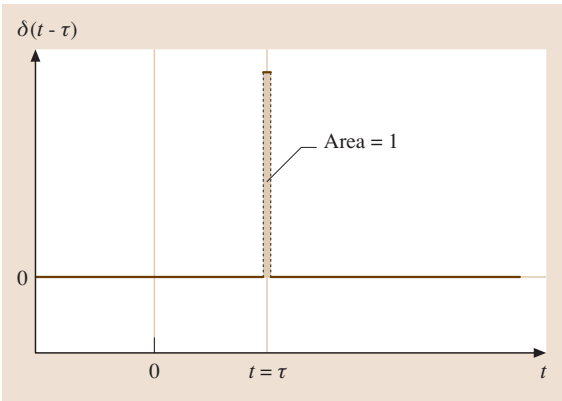


Fig. 1.17 The unit impulse occurring at $t = \tau$

have

$$\mathbf{L}[\delta(t)] \overset{\text{Notation}}{=} \Delta(s) = \lim_{t_1 \to 0} \left(\frac{1 - \mathrm{e}^{-st_1}}{st_1} \right) \overset{\text{L'Hôspital's rule}}{=} \lim_{t_1 \to 0} \left(\frac{s\,\mathrm{e}^{-st_1}}{s} \right) = 1 \,. \tag{1.68}$$

If the unit impulse occurs at some time $t = \tau$ (Fig. 1.17) it is represented by $\delta(t - \tau)$, and

$$\mathbf{L}[\delta(t - \tau)] = \mathrm{e}^{-\tau s} \,. \tag{1.69}$$

This signal has the property $\delta(t - \tau) = 0$ for $t \neq \tau$, $\delta(t - \tau) = \infty$ for $t = \tau$, and $\int_{-\infty}^{\infty} \delta(t - \tau)\,\mathrm{d}t = 1$. It also has the filtering property,

$$\int_{-\infty}^{\infty} f(\tau)\delta(t - \tau)\,\mathrm{d}\tau = f(t) \,. \tag{1.70}$$

1.3.3 Laplace Transform of Derivatives and Integrals

Since engineering systems are generally modeled by differential equations of various orders, we need to have knowledge of the Laplace transform of derivatives of different orders. In other occasions, the system may be described by an equation that contains not only derivatives, but also integrals; for instance, a circuit involving a resistor, an inductor, and a capacitor (RLC circuit) [1.3]. We will also present a systematic approach for solving initial-value problems.

Theorem 1.5: Laplace transform of derivatives
If $F(s) = \mathbf{L}[f(t)]$, then

$$\mathbf{L}[\dot{f}(t)] = sF(s) - f(0) \tag{1.71}$$

and

$$\mathbf{L}[\ddot{f}(t)] = s^2 F(s) - sf(0) - \dot{f}(0) \,. \tag{1.72}$$

In general,

$$\mathbf{L}[f^{(n)}(t)] = s^n F(s) - s^{n-1} f(0) - s^{n-2} \dot{f}(0) - \cdots - f^{(n-1)}(0) \,. \tag{1.73}$$

Theorem 1.6: Laplace transform of integrals
If $F(s) = \mathbf{L}[f(t)]$, then

$$\mathbf{L}\left[\int_0^t f(\tau)\,\mathrm{d}\tau \right] = \frac{1}{s} F(s) \,. \tag{1.74}$$

Alternatively,

$$\mathbf{L}^{-1}\left[\frac{1}{s} F(s) \right] = \int_0^t f(\tau)\,\mathrm{d}\tau \,. \tag{1.75}$$

Solving Initial-Value Problems. The role of the Laplace transforms of derivatives and integrals of time-varying functions is most significant when solving an initial-value problem. Schematically, the solution method is as in Fig. 1.18.

Example 1.24: Second-order IVP
Solve $\ddot{x} + 2\dot{x} + x = 0$, $x(0) = 1$, $\dot{x}(0) = 1$.

Solution. Laplace transformation results in

$$[s^2 X(s) - sx(0) - \dot{x}(0)] + 2[sX(s) - x(0)] + X(s) = 0 \overset{\text{Solve for X(s)}}{\Rightarrow} X(s) = \frac{s + 3}{(s + 1)^2} \,.$$

Before inversion, we rewrite this last expression as

$$X(s) = \frac{s + 3}{(s + 1)^2} = \frac{(s + 1) + 2}{(s + 1)^2} = \frac{1}{s + 1} + \frac{2}{(s + 1)^2} \,.$$

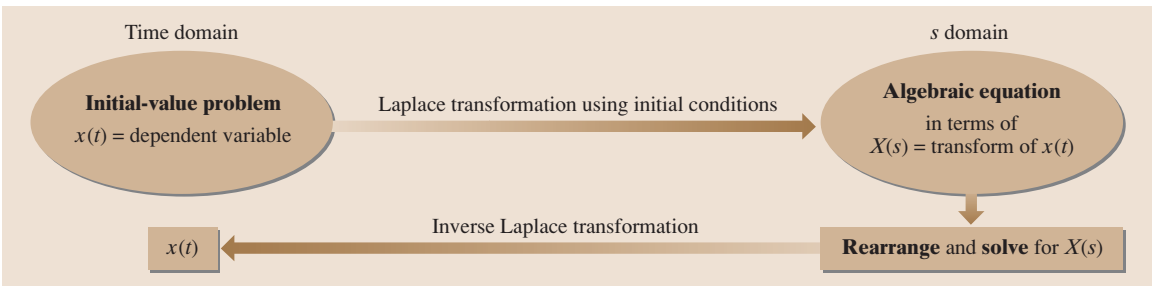


Fig. 1.18 The solution method for initial-value problems

Finally,

$$x(t) = \mathbf{L}^{-1}[X(s)] = \mathrm{e}^{-t} + 2t\,\mathrm{e}^{-t} = (2t+1)\mathrm{e}^{-t}\,.$$

1.3.4 Inverse Laplace Transformation

Inverse Laplace transformation clearly plays a vital role in completing the procedure for solving differential equations. In this section we will learn a systematic technique, using partial fractions, to treat a wide range of inverse Laplace transforms. We will also introduce the convolution method, which is quite important from a physical standpoint.

Partial Fractions Method

When solving an ODE in terms of $x(t)$ through Laplace transformation, the very last step involves finding $\mathbf{L}^{-1}[X(s)]$. And we almost always find ourselves looking for the inverse Laplace transform of functions in the form of

$$X(s) = \frac{N(s)}{D(s)} = \frac{\text{Polynomial of degree } m}{\text{Polynomial of degree } n},\ m < n\,. \tag{1.76}$$

The case $m \geq n$ is purely mathematical and does not occur in engineering analysis. The idea behind the partial fractions method is simple: express $X(s) = N(s)/D(s)$ as a suitable sum of fractions, and find $\mathbf{L}^{-1}$ of each fraction accordingly. So, it all boils down to how to break the original fraction into partial fractions, and this depends on the nature of the poles of $X(s)$ – the roots of $D(s)$. In the most general sense, these roots can be real or complex. For instance, $D(s) = s^3 - s^2 - 2s$ has roots $s = 0, -1, 2$, all real, so we express $D(s)$ as $D(s) = s(s+1)(s-2)$. On the other hand, the roots of $D(s) = s^3 + 2s^2 + 2s$ are $s = 0, -1 \pm \mathrm{i}$. In this case, instead of writing $D(s) = s(s+1+\mathrm{i})(s+1-\mathrm{i})$ we write $D(s) = s(s^2+2s+2)$. In other words, the quadratic polynomial with complex roots remains intact. Any second-degree polynomial with complex roots is called an irreducible polynomial.

Structure of Partial Fractions. There are four possible cases that arise in the process of finding $\mathbf{L}^{-1}[X(s)]$. We must first identify the factors of $D(s)$, which could be linear, repeated linear, irreducible polynomial, or repeated irreducible polynomial. Then, each identified factor will have its own fraction (or fractions) associated with them. The details follow.

Case 1: Linear Factor $s - p_i$. Each typical linear factor $s - p_i$ of $D(s)$ is associated with a fraction in the form

$$\frac{A}{s-p_i}\,,$$

where $A = \text{const.}$ is to be determined appropriately. We note that $s = p_i$ is called a simple pole of $X(s)$.

Example 1.25: Linear factors
Find $\mathbf{L}^{-1}[X(s)]$ where

$$X(s) = \frac{s+1}{(s+3)(s+4)}\,.$$

Solution. The poles of $X(s)$ are $p_1 = -3$, $p_2 = -4$. Since there are two linear factors, there will be two fractions. We write

$$X(s) = \frac{s+1}{(s+3)(s+4)} = \frac{A_1}{s+3} + \frac{A_2}{s+4} = \frac{A_1(s+4)+A_2(s+3)}{(s+3)(s+4)} \overset{\text{Collect like terms}}{=} \frac{(A_1+A_2)s + 4A_1 + 3A_2}{(s+3)(s+4)}\,,$$

$$A_1,\ A_2 = \text{const.}$$

The denominators of the original and the final fractions are identical (by design), so we force their respective numerators to be identical, that is,

$$s + 1 \equiv (A_1 + A_2)s + 4A_1 + 3A_2\,.$$

But this identity holds only if the coefficients of like powers of s on both sides are the same. So, we have

$$\begin{aligned} &\text{Coefficient of } s\text{:} && 1 = A_1 + A_2 \\ &\text{Constant term:} && 1 = 4A_1 + 3A_2\,. \end{aligned} \quad \overset{\text{Solve}}{\Rightarrow} \quad \begin{aligned} A_1 &= -2 \\ A_2 &= 3 \end{aligned}$$

Insert these into the partial fractions, and perform a term-by-term inverse Laplace transformation, to obtain

$$X(s) = \frac{-2}{s+3} + \frac{3}{s+4} \quad \overset{\mathbf{L}^{-1}}{\Rightarrow}$$

$$x(t) \overset{\text{Linearity}}{=} -2\mathbf{L}^{-1}\left(\frac{1}{s+3}\right) + 3\mathbf{L}^{-1}\left(\frac{1}{s+4}\right) = -2\mathrm{e}^{-3t} + 3\mathrm{e}^{-4t}\,.$$

Case 2: Repeated Linear Factor $(s - p_i)^k$. If a root of $D(s)$, say, p_i, happens to have multiplicity k, then $D(s)$ contains the factor $(s - p_i)^k$. This factor is then associated with partial fractions

$$\frac{A_k}{(s-p_i)^k} + \frac{A_{k-1}}{(s-p_i)^{k-1}} + \cdots + \frac{A_2}{(s-p_i)^2} + \frac{A_1}{s-p_i}\,,$$

where the constants $A_k, \cdots, A_1$ are determined as in case 1. As an example, we first write

$$X(s) = \frac{4s+7}{(s+1)^2(s+4)} = \underbrace{\frac{A_2}{(s+1)^2}}_{\text{Double pole; case 2}} + \frac{A_1}{s+1} + \underbrace{\frac{A_3}{s+4}}_{\text{Simple pole; case 1}}$$

and then proceed as before to determine the constants.

Case 3: Irreducible Polynomial s^2+as+b. This occurs when $X(s)$ has a pair of complex-conjugate poles. Each irreducible polynomial is associated with a single fraction in the form

$$\frac{Bs+C}{s^2+as+b},$$

where the constants B and C are found as before. Before taking the inverse Laplace transform, we must first complete the square in the irreducible polynomial, that is, $s^2+as+b=(s+\sigma)^2+\omega^2$. Then, at some point, we need to determine

$$\mathbf{L}^{-1}\left[\frac{Bs+C}{(s+\sigma)^2+\omega^2}\right].$$

The key is to split the fraction in terms of the two expressions

$$\frac{\omega}{(s+\sigma)^2+\omega^2} \quad \text{and} \quad \frac{s+\sigma}{(s+\sigma)^2+\omega^2}$$

so that we can ultimately use the relations [see Table 1.2]

$$\mathbf{L}^{-1}\left[\frac{\omega}{(s+\sigma)^2+\omega^2}\right] = \mathrm{e}^{-\sigma t}\sin\omega t\,,$$
$$\mathbf{L}^{-1}\left[\frac{s+\sigma}{(s+\sigma)^2+\omega^2}\right] = \mathrm{e}^{-\sigma t}\cos\omega t\,. \qquad (1.77)$$

Case 4: Repeated Irreducible Polynomial $(s^2+as+b)^k$. The fractions are formed as

$$\frac{B_k s+C_k}{(s^2+as+b)^k}+\cdots+\frac{B_2 s+C_2}{(s^2+as+b)^2}+\frac{B_1 s+C_1}{s^2+as+b}.$$

Convolution Method

In systems analysis, the problem of determining the time history of a function often comes down to $\mathbf{L}^{-1}[G(s)H(s)]$, where the inverse Laplace transforms of $G(s)$ and $H(s)$ are known. The convolution method allows us to determine $\mathbf{L}^{-1}[G(s)H(s)]$ using knowledge of $g(t)$ and $h(t)$.

Notation: $\mathbf{L}^{-1}[G(s)H(s)] \overset{\text{Notation}}{=} (g*h)(t)$ is read "*convolution* of g and h"

Theorem 1.7: Convolution. Let $G(s)=\mathbf{L}[g(t)]$, $H(s)=\mathbf{L}[h(t)]$, and $F(s)=G(s)H(s)$. Then,

$$\mathbf{L}^{-1}[F(s)] = f(t) = (g*h)(t) = \int_0^t g(\tau)h(t-\tau)\,\mathrm{d}\tau = \int_0^t h(\tau)g(t-\tau)\,\mathrm{d}\tau = (h*g)(t)\,:$$

The result clearly indicates that the convolution of two functions is *symmetric*.

Example 1.26: Convolution
Find $\mathbf{L}^{-1}\left[\frac{1}{s^2(s+2)}\right]$.

Solution. Write $F(s)=\frac{1}{s^2}\cdot\frac{1}{s+2}=G(s)H(s)$ and pick $G(s)=1/s^2$, $H(s)=1/(s+2)$ so that $g(t)=t$, $h(t)=\mathrm{e}^{-2t}$. Then

$$f(t)=(g*h)(t)=\int_0^t \tau\cdot\mathrm{e}^{-2(t-\tau)}\,\mathrm{d}\tau$$
$$\overset{\text{Integration by parts}}{=}\left[\tau\cdot\frac{1}{2}\mathrm{e}^{-2(t-\tau)}\right]_{\tau=0}^{t} - \frac{1}{2}\int_0^t \mathrm{e}^{-2(t-\tau)}\,\mathrm{d}\tau$$
$$=\frac{1}{2}t-\frac{1}{4}(1-\mathrm{e}^{-2t})=\frac{1}{4}(\mathrm{e}^{-2t}+2t-1)\,.$$

1.3.5 Periodic Functions

Physical systems are often subjected to external disturbances that exhibit repeated behavior over long periods

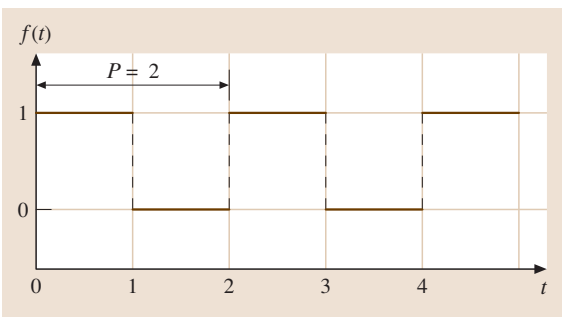


Fig. 1.19 Periodic function of Example 1.27

of time. A function $f(t)$ is called periodic with *period* $P > 0$ if it is defined for all $t > 0$, and $f(t+P) = f(t)$ for all $t > 0$.

It can then be shown [1.1] that the Laplace transform of this function is

$$F(s) = \frac{1}{1 - \mathrm{e}^{-Ps}} \int_0^P \mathrm{e}^{-st} f(t)\,\mathrm{d}t \,. \tag{1.78}$$

Example 1.27: Periodic signal
Find the Laplace transform of the periodic function in Fig. 1.19.

Solution. It is evident that the period is $P = 2$. With this, the integral in (1.78) is

$$\int_0^2 \mathrm{e}^{-st} f(t)\,\mathrm{d}t$$

$$\underset{f(t)=0 \text{ for } 1<t<2}{\overset{f(t)=1 \text{ for } 0<t<1}{=}} \int_0^1 \mathrm{e}^{-st}\,\mathrm{d}t = (1 - \mathrm{e}^{-s})/s \,.$$

Then, by (1.78), $F(s) = (1 - \mathrm{e}^{-s})/(s(1 - \mathrm{e}^{-2s}))$. Noting that $1 - \mathrm{e}^{-2s} = 1 - (\mathrm{e}^{-s})^2 = (1 - \mathrm{e}^{-s})(1 + \mathrm{e}^{-s})$, the above expression reduces to $F(s) = 1/(s(1 + \mathrm{e}^{-s}))$.

1.4 Fourier Analysis

This section will focus on the concepts of Fourier series and transformation and their properties. The idea of Fourier series is based on representing periodic functions in terms of series of sine and cosine components whose periods are integral multiples of each other. The Fourier transform is an operation that maps a function defined in the time domain into one in the frequency domain; its extension leads to the familiar Laplace transformation.

1.4.1 Fourier Series

Let $f(x)$ be periodic with period $P > 0$, that is, $f(x)$ is defined for all real x, and $f(x+P) = f(x)$ for all x. Then, assuming that $f(x)$ has a Fourier series representation, it is in the form

$$\frac{1}{2}a_0 + \sum_{n=1}^{\infty} \left(a_n \cos\frac{2n\pi x}{P} + b_n \sin\frac{2n\pi x}{P} \right) , \tag{1.79}$$

where the constants are generated by the *Euler–Fourier formulas*, as [1.1, 7]

$$a_0 = \frac{2}{P} \int_{-P/2}^{P/2} f(x)\,\mathrm{d}x \,, \tag{1.80}$$

$$a_n = \frac{2}{P} \int_{-P/2}^{P/2} f(x) \cos\frac{2n\pi x}{P}\,\mathrm{d}x \,,$$

$$n = 1, 2, 3, \cdots \,, \tag{1.81}$$

$$b_n = \frac{2}{P} \int_{-P/2}^{P/2} f(x) \sin\frac{2n\pi x}{P}\,\mathrm{d}x \,,$$

$$n = 1, 2, 3, \cdots \,. \tag{1.82}$$

In (1.79), each of the sinusoidal components with different frequencies is called a harmonic, with *amplitude* a_n or b_n. The harmonics basically describe the variations of $f(x)$ about its average value of $\frac{1}{2}a_0$. In certain cases, only a few harmonics may be needed to represent $f(x)$ with reasonable accuracy. On the other hand, there are situations where many of them may be required. At the points of discontinuity of $f(x)$ the partial sums assume the average value of the left- and right-hand limits.

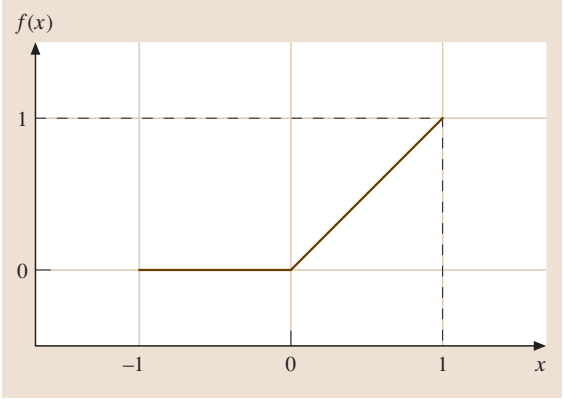


Fig. 1.20 A periodic function with period $P = 2$

Example 1.28: Fourier series

Find the Fourier series representation of the periodic function whose description in one period is shown in Fig. 1.20.

Solution

By (1.80), we have $a_0 = \int_{-1}^{1} f(x)\,\mathrm{d}x = \int_{0}^{1} x\,\mathrm{d}x = 1/2$. For $n = 1, 2, 3, \cdots$, (1.81) and (1.82) yield

$$
\begin{aligned}
a_n &= \int_{-1}^{1} f(x)\cos n\pi x\,\mathrm{d}x = \int_{0}^{1} x\cos n\pi x\,\mathrm{d}x \\
&= \frac{1}{(n\pi)^2}(\cos n\pi - 1) \\
&= \frac{1}{(n\pi)^2}[(-1)^n - 1] = \begin{cases} 0 & \text{if} \quad n = \text{even} \\ -\frac{2}{(n\pi)^2} & \text{if} \quad n = \text{odd}\,, \end{cases} \\
b_n &= \int_{-1}^{1} f(x)\sin n\pi x\,\mathrm{d}x = \int_{0}^{1} x\sin n\pi x\,\mathrm{d}x \\
&= -\frac{1}{n\pi}[x\cos n\pi]_0^1 + \frac{1}{(n\pi)^2}[\sin n\pi x]_0^1 \\
&= \frac{1}{n\pi}(-1)^{n+1}\,.
\end{aligned}
$$

Equation (1.79) gives the Fourier series of $f(x)$, as

$$
\begin{aligned}
&\frac{1}{4} - \frac{2}{\pi^2}\cos\pi x + \frac{1}{\pi}\sin\pi x - \frac{1}{2\pi}\sin 2\pi x \\
&- \frac{2}{9\pi^2}\cos 3\pi x + \frac{1}{3\pi}\sin 3\pi x + \cdots \\
&\overset{\text{Collect terms}}{=} \frac{1}{4} - \frac{2}{\pi^2}\left(\cos\pi x + \frac{1}{9}\cos 3\pi x + \cdots\right) \\
&+ \frac{1}{\pi}\left(\sin\pi x - \frac{1}{2}\sin 2\pi x + \frac{1}{3}\sin 3\pi x - \cdots\right)\,.
\end{aligned}
$$

The third and ninth partial sums, together with the original function, are shown in Fig. 1.21. As mentioned earlier, at the points of discontinuity of $f(x)$ the partial sums assume the average value of the left- and right-hand limits, that is, 1/2.

1.4.2 Fourier Transformation

In Sect. 1.3 the notation $F(s)$ was used to represent the Laplace transform of $f(t)$, i. e., $F(s) = \mathbf{L}[f(t)]$. Similarly, $\hat{f}(\omega)$ is used to denote the Fourier transform of $f(t)$, that is, $\hat{f}(\omega) = \mathbf{F}[f(t)]$. Since ω is complex in general, $\hat{f}(\omega)$ is expected to be complex-valued as well. With this, we then write $f(t) = \mathbf{F}^{-1}[\hat{f}(\omega)]$ describing the inverse Fourier transform of $\hat{f}(\omega)$. We define the Fourier transform pair as

$$
\hat{f}(\omega) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f(\tau)\,\mathrm{e}^{-\mathrm{i}\omega\tau}\,\mathrm{d}\tau\,, \tag{1.83}
$$

$$
f(t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \hat{f}(\omega)\,\mathrm{e}^{\mathrm{i}\omega t}\,\mathrm{d}\omega\,. \tag{1.84}
$$

Fourier transformation can be thought of as a *mapping* that assigns to a given function of time t an integral function of frequency ω. In general, any trans-

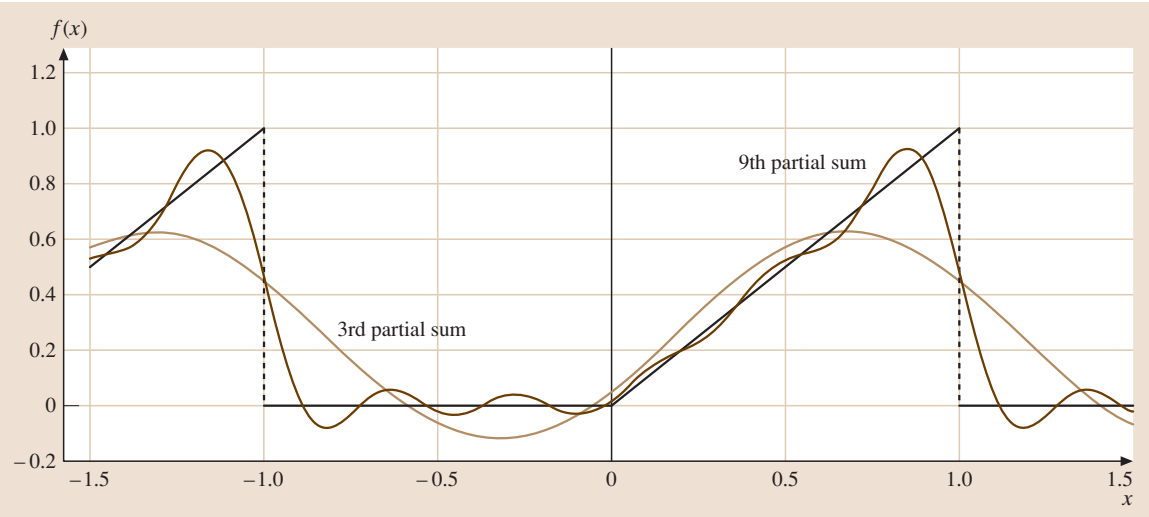


Fig. 1.21 Example 1.28

Table 1.3 Fourier transform pairs

No.	$f(t)$	$\hat{f}(\omega)$
1	$\begin{cases} 1 & -b < t < b \\ 0 & \text{otherwise} \end{cases}$	$\sqrt{\frac{2}{\pi}}\,\frac{\sin b\omega}{\omega}$
2	$\begin{cases} 1 & b_1 < t < b_2 \\ 0 & \text{otherwise} \end{cases}$	$\frac{1}{\sqrt{2\pi}}\,\frac{e^{-ib_1\omega}-e^{-ib_2\omega}}{i\omega}$
3	$\begin{cases} e^{-at} & t > 0 \\ 0 & \text{otherwise} \end{cases}, a > 0$	$\frac{1}{\sqrt{2\pi}}\,\frac{1}{a+i\omega}$
4	$\begin{cases} e^{at} & t < 0 \\ 0 & \text{otherwise} \end{cases}, a > 0$	$\frac{1}{\sqrt{2\pi}}\,\frac{1}{a-i\omega}$
5	$\begin{cases} e^{at} & b_1 < t < b_2 \\ 0 & \text{otherwise} \end{cases}$	$\frac{1}{\sqrt{2\pi}}\,\frac{e^{(a-i\omega)b_2}-e^{(a-i\omega)b_1}}{a-i\omega}$
6	$e^{-a\lvert t\rvert}, a > 0$	$\sqrt{\frac{2}{\pi}}\,\frac{a}{\omega^2+a^2}$
7	$\begin{cases} -e^{-at} & t < 0 \\ e^{at} & t > 0 \end{cases}, a < 0$	$\sqrt{\frac{2}{\pi}}\,\frac{-i\omega}{\omega^2+a^2}$
8	$\begin{cases} e^{iat} & -b < t < b \\ 0 & \text{otherwise} \end{cases}$	$\sqrt{\frac{2}{\pi}}\,\frac{\sin(\omega-a)b}{\omega-a}$
9	$\begin{cases} e^{iat} & b_1 < t < b_2 \\ 0 & \text{otherwise} \end{cases}$	$\frac{i}{\sqrt{2\pi}}\,\frac{e^{i(a-\omega)b_1}-e^{i(a-\omega)b_2}}{a-\omega}$
10	$\frac{1}{a^2+t^2}, a > 0$	$\sqrt{\frac{\pi}{2}}\,\frac{e^{-a\lvert\omega\rvert}}{a}$
11	$\begin{cases} t & 0 < t < b \\ 0 & \text{otherwise} \end{cases}$	$\frac{1}{\sqrt{2\pi}}\,\frac{-1+e^{-ib\omega}(1+ib\omega)}{\omega^2}$
12	$\begin{cases} t & 0 < t < b \\ 2t-b & b < t < 2b \\ 0 & \text{otherwise} \end{cases}$	$\frac{1}{\sqrt{2\pi}}\,\frac{-1+2e^{ib\omega}-e^{-2ib\omega}}{\omega^2}$
13	$e^{-at^2}, a > 0$	$\frac{1}{\sqrt{2a}}\,e^{-\omega^2/(4a)}$
14	$e^{-t^2/(4a)}, a > 0$	$\sqrt{2a}\,e^{-a\omega^2}$

formation with this type of property is known as an *integral transformation*. The obvious similarities between the Laplace and Fourier transforms are credited to the Laplace transformation being an integral transformation itself. Fourier transforms of several functions are listed in Table 1.3.

Example 1.29: Fourier transform
Find the Fourier transform of

$$f(t) = \begin{cases} 0 & \text{if} \quad t < 0 \\ e^{-at} & \text{if} \quad t > 0 \end{cases} (a > 0)\,.$$

Solution. By (1.83),

$$\begin{aligned} \hat{f}(\omega) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(\tau)\, e^{-i\omega\tau}\, d\tau \\ &= \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} e^{-a\tau}\, e^{-i\omega\tau}\, d\tau \\ &= \frac{1}{\sqrt{2\pi}} \frac{-1}{a+i\omega} \Big[e^{-(a+i\omega)\tau} \Big]_0^{\infty} \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{a+i\omega}\,. \end{aligned}$$

Using $\hat{f}(\omega)$ above in (1.84), we find

$$\begin{aligned} f(t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \frac{1}{a+i\omega}\, e^{i\omega t}\, d\omega \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{1}{a+i\omega}\, e^{i\omega t}\, d\omega\,. \end{aligned}$$

This is known as the complex Fourier integral representation of the function under consideration.

1.5 Linear Algebra

In this section we present the fundamentals of linear algebra, specifically, vectors and matrices, and their relation to linear systems of algebraic and differential equations. The methods of linear algebra are mainly useful in the treatment of systems of equations that are heavily coupled, that is, when a large number of equations in the system involve many of the unknown variables. In these cases, techniques such as direct substitution and elimination are no longer suitable due to their lack of computational efficiency. We focus on algebraic systems first, then extend the ideas to systems of differential equations.

1.5.1 Vectors and Matrices

An n-dimensional vector $\boldsymbol{v}$ is an ordered set of n scalars, and is written as $\boldsymbol{v} = (v_1, v_2, \cdots, v_n)$. Each v_i $(i = 1, 2, \cdots, n)$ is called a component of the vector $\boldsymbol{v}$. In a general sense, an n-dimensional vector helps us locate a point in an n-dimensional space, regardless of the nature of its components or perhaps what physical quantity each may represent.

Matrices

A collection of numbers (real or complex) or possibly functions, arranged in a rectangular array and enclosed by brackets, is referred to as a matrix. Each of the elements in a matrix is called an entry (or element) of the matrix. The horizontal and vertical lines are referred to as rows and columns of the matrix, respectively. A matrix is called a row vector if it consists of one row only, and a column vector if it has one column only. The number of rows and columns of a matrix determine the size of that matrix. If a matrix $\mathbf{A}$ has m rows and n columns, then it is said to be of size $m \times n$. If the number of rows and columns are the same, we speak of a square matrix, otherwise, a rectangular matrix. We denote matrices by bold-faced capital letters, such as $\mathbf{A}$. The abbreviated form of an $m \times n$ matrix is

$$\mathbf{A} = [a_{ij}]_{m\times n}\ ,\ i = 1, 2, \cdots, m\ ,\ j = 1, 2, \cdots, n\ ,$$

where a_{ij} is known as the (i, j) entry of $\mathbf{A}$, located at the intersection of the ith row and the jth column of $\mathbf{A}$ so that a_{12}, for instance, occupies the entry at which the first row and the second column meet. In the event that $\mathbf{A}$ is a square matrix $(m = n)$, the elements $a_{11}, a_{22}, \cdots, a_{nn}$ are referred to as the *diagonal entries* of $\mathbf{A}$. These diagonal elements form the main diagonal of $\mathbf{A}$. The diagonal directly below the main is known as the *subdiagonal* and the one above the main is called the *superdiagonal*. Two matrices $\mathbf{A} = [a_{ij}]$ and $\mathbf{B} = [b_{ij}]$ are said to be equal if they have the same size, and the same entries in the respective locations. If some rows or columns (or possibly both) of $\mathbf{A}$ are deleted, the outcome is a submatrix of $\mathbf{A}$. If no rows or columns of $\mathbf{A}$ are omitted, we have $\mathbf{A}$ as a submatrix of itself. Submatrices play important roles in such areas of matrix analysis as determinants and rank.

Matrix Operations

Matrices of the same size can be added. The result, or the sum, is a matrix of the same size. If $\mathbf{A} = [a_{ij}]$ and $\mathbf{B} = [b_{ij}]$ are $m \times n$, their sum $\mathbf{C} = [c_{ij}]$ is also $m \times n$. Matrix addition is performed entry-wise, that is, the entry of $\mathbf{C}$ in the (i, j) slot is the sum of the entries of $\mathbf{A}$ and $\mathbf{B}$ in that same slot. The $m \times n$ zero matrix, denoted by $\mathbf{0}_{m\times n}$, is an $m \times n$ matrix all of whose entries are zero. If $\mathbf{A} = [a_{ij}]$ is $m \times n$ and k is a scalar, then $k\mathbf{A}$ is an $m \times n$ matrix whose entries are those of $\mathbf{A}$ multiplied by k in every slot, that is, $k\mathbf{A} = [ka_{ij}]_{m\times n}$. Let $\mathbf{A} = [a_{ij}]_{m\times n}$ and $\mathbf{B} = [b_{ij}]_{n\times p}$. It is important to note that the number of columns of $\mathbf{A}$ is n, which is equal to the number of rows of $\mathbf{B}$. Then, their product $\mathbf{C} = \mathbf{AB}$ is $m \times p$ whose entries are obtained as

$$c_{ij} = \sum_{k=1}^{n} a_{ik}b_{kj}\ ,\ i = 1, 2, \cdots, m\ ,\ j = 1, 2, \ldots, p\ . \tag{1.85}$$

This is shown schematically in Fig. 1.22. If the number of columns of $\mathbf{A}$ does not match the number of rows of

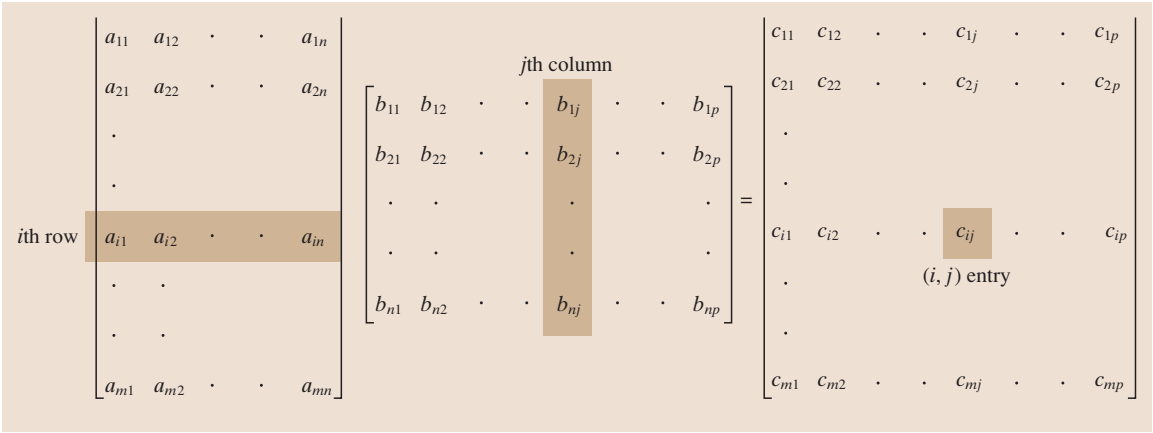


Fig. 1.22 Construction of the matrix product $\mathbf{AB} = \mathbf{C}$

B, the product is undefined. If the product is defined, then to get the (i, j) entry of **C**, we proceed as follows: the ith row of **A** is clearly a $1\times n$ vector. The jth column of **B** is an $n\times 1$ vector, hence these two vectors have the same number of components, n. In these two vectors, multiply the first components, the second components, etc., up to the nth components. Then add the individual products together. The result is c_{ij}.

Example 1.30: Matrix Multiplication
Find

$$\mathbf{AB} = \begin{pmatrix} 1 & -2 & 3 \\ 0 & 1 & 4 \end{pmatrix}_{2\times 3} \begin{pmatrix} -2 & -1 & 4 \\ 1 & 2 & 0 \\ 3 & 5 & 1 \end{pmatrix}_{3\times 3} .$$

Solution. We first note that the operation is valid because **A** has three columns and **B** has three rows. And, **AB** will be 2×3. Following the strategy outlined above, we find the product as

$$\mathbf{AB} = \begin{pmatrix} 1\cdot(-2)+(-2)\cdot 1+3\cdot 3 & 1\cdot(-1)+(-2)\cdot 2+3\cdot 5 & 1\cdot 4+(-2)\cdot 0+3\cdot 1 \\ 0\cdot(-2)+1\cdot 1+4\cdot 3 & 0\cdot(-1)+1\cdot 2+4\cdot 5 & 0\cdot 4+1\cdot 0+4\cdot 1 \end{pmatrix} = \begin{pmatrix} 5 & 10 & 7 \\ 13 & 22 & 4 \end{pmatrix}_{2\times 3} .$$

Matrix Transpose

Given an $m\times n$ matrix **A**, its transpose, denoted by $\mathbf{A}^{\mathrm{T}}$, is an $n\times m$ matrix with the property that its first row is the first column of **A**, its second row is the second column of **A**, and so on. Given that all matrix operations are valid,

$$(\mathbf{A}+\mathbf{B})^{\mathrm{T}} = \mathbf{A}^{\mathrm{T}} + \mathbf{B}^{\mathrm{T}} \tag{1.86}$$

$$(k\mathbf{A})^{\mathrm{T}} = k\mathbf{A}^{\mathrm{T}}, \quad \text{scalar } k \tag{1.87}$$

$$(\mathbf{AB})^{\mathrm{T}} = \mathbf{B}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}} . \tag{1.88}$$

Special Matrices

A square matrix **A** is symmetric if $\mathbf{A}^{\mathrm{T}} = \mathbf{A}$ and skew-symmetric if $\mathbf{A}^{\mathrm{T}} = -\mathbf{A}$. A square matrix $\mathbf{A}_{n\times n} = [a_{ij}]$ is called upper-triangular if $a_{ij} = 0$ for all $i > j$, that is, every entry below the main diagonal is zero, lower-triangular if $a_{ij} = 0$ for all $i < j$, that is, all elements above the main diagonal are zeros, and diagonal if $a_{ij} = 0$ for all $i \neq j$. The $n\times n$ identity matrix is a diagonal matrix whose diagonal entries are all equal to 1, and is denoted by **I**.

Example 1.31: Special matrices
Matrices **U**, **L**, and **D** are upper triangular, lower triangular, and diagonal, respectively:

$$\mathbf{U} = \begin{pmatrix} -2 & 1 & 2 \\ 0 & 5 & 0 \\ 0 & 0 & 3 \end{pmatrix} , \quad \mathbf{L} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 0 & 0 \\ 4 & 7 & -1 \end{pmatrix} ,$$

$$\mathbf{D} = \begin{pmatrix} 3 & 0 & 0 \\ 0 & -4 & 0 \\ 0 & 0 & 1 \end{pmatrix} .$$

Note that in **U** and **L** zeros are allowed along the main diagonal. In fact, the main diagonal may consist of all zeros. On the other hand, **D** may have one or more zero diagonal elements, as long as they are not all zeros. In the event that all entries of an $n\times n$ matrix are zeros, it is called the $n\times n$ *zero matrix* $\mathbf{0}_{n\times n}$.

Determinant

The determinant of a square matrix $\mathbf{A} = [a_{ij}]_{n\times n}$ is a real scalar denoted by $|\mathbf{A}|$ or $\det(\mathbf{A})$. For the most trivial case of $n = 1$, $\mathbf{A} = [a_{11}]$, and we define the determinant simply as $|\mathbf{A}| = a_{11}$. For $n \geq 2$, the determinant is defined as

using the i-th row

$$|\mathbf{A}| = \sum_{k=1}^{n} a_{ik}(-1)^{i+k} M_{ik} , \quad i = 1, 2, \cdots, n \tag{1.89}$$

or

using the j-th column

$$|\mathbf{A}| = \sum_{k=1}^{n} a_{kj}(-1)^{k+j} M_{kj} , \quad j = 1, 2, \cdots, n \tag{1.90}$$

Here M_{ik} is the minor of the entry a_{ik}, defined as the determinant of the $(n-1)\times(n-1)$ submatrix of **A** obtained by deleting the ith row and the kth column of **A**. The quantity $(-1)^{i+k}M_{ik}$ is known as the cofactor of a_{ik} and is denoted by C_{ik}. Also note that $(-1)^{i+k}$ is responsible for whether a term is multiplied by $+1$ or -1. Equations (1.89) and (1.90) suggest that the determinant of a square matrix can be calculated using any row or any column of the matrix. However, for all practical purposes, it is wise to use the row (or column) containing the most number of zeros, or if none, the one with the smallest entries. A square matrix with a nonzero determinant is known as a nonsingular matrix. Otherwise, it is called singular. The rank of any matrix **A**, denoted by $\mathrm{rank}(\mathbf{A})$, is the size of the largest nonsingular submatrix of **A**. If $|\mathbf{A}_{n\times n}| = 0$, we conclude that $\mathrm{rank}\,(\mathbf{A}) < n$.

Example 1.32: 3×3 determinant
Find the determinant of

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & -3 \\ 4 & -1 & 1 \\ 2 & 0 & 1 \end{pmatrix} .$$

Solution. We will use the third row because it happens to contain a zero. Following (1.89),

$$\begin{aligned} |\mathbf{A}| &= 2\cdot(-1)^{3+1}M_{31} + 0 + 1\cdot(-1)^{3+3}M_{33} \\ &= 2\begin{vmatrix} 2 & -3 \\ -1 & 1 \end{vmatrix} + \begin{vmatrix} 1 & 2 \\ 4 & -1 \end{vmatrix} \\ &= 2(2-3) + (-1-8) = -11 . \end{aligned}$$

Properties of Determinant. The determinant of a matrix possesses a number of important properties, some of which are listed below [1.1]:

- A square matrix $\mathbf{A}$ and its transpose have the same determinant, that is, $|\mathbf{A}| = |\mathbf{A}^{\mathrm{T}}|$.
- The determinant of diagonal, upper-triangular and lower-triangular matrices is the product of the diagonal entries.
- If an entire row (or column) of a square matrix $\mathbf{A}$ is zero, then $|\mathbf{A}| = 0$.
- If $\mathbf{A}$ is $n\times n$ and k is scalar, then $|k\mathbf{A}| = k^n\,|\mathbf{A}|$.
- If any two rows (or columns) of $\mathbf{A}$ are interchanged, the determinant of the resulting matrix is $-|\mathbf{A}|$.
- The determinant of the product of two matrices obeys $|\mathbf{AB}| = |\mathbf{A}|\,|\mathbf{B}|$.
- Any square matrix with any number of linearly dependent rows (or columns) is singular.

Determinant of Block Matrices. We define a block-diagonal matrix as a square matrix partitioned such that its diagonal elements are square matrices, while all other elements are zeros; see Fig. 1.23a. Similarly, a block-triangular matrix is a square matrix partitioned so that its diagonal elements are square blocks, while all entries either above or below this main block diagonal are zeros; see Fig. 1.23b,c.

Many properties of these special block matrices are basically extensions of those of diagonal and triangular matrices. In particular, the determinant of each of these matrices is equal to the product of the individual determinants of the blocks along the main diagonal. Consequently, a block diagonal (or triangular) matrix is singular if and only if one of the blocks along the main diagonal is singular.

Inverse of a Matrix. Given a square matrix $\mathbf{A}_{n\times n}$, its inverse is denoted by $\mathbf{A}^{-1}$ with the property that

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{I} = \mathbf{A}^{-1}\mathbf{A} , \tag{1.91}$$

where $\mathbf{I}$ denotes the $n\times n$ identity matrix. If $\mathbf{A}^{-1}$ exists, then it is unique. A square matrix has an inverse if and only if it is nonsingular. Equivalently, $\mathbf{A}_{n\times n}$ has an inverse if and only if rank $(\mathbf{A}) = n$. A square matrix with an inverse is called invertible. An immediate application of the inverse is in the solution process of a linear system $\mathbf{A}\boldsymbol{x} = \boldsymbol{b}$. Multiplying this equation from the left, known as premultiplication, by $\mathbf{A}^{-1}$, yields

$$\begin{aligned} & \mathbf{A}^{-1}(\mathbf{A}\boldsymbol{x}) = \mathbf{A}^{-1}\boldsymbol{b} \\ \Rightarrow\ & (\mathbf{A}^{-1}\mathbf{A})\boldsymbol{x} = \mathbf{A}^{-1}\boldsymbol{b} \\ \Rightarrow\ & \mathbf{I}\boldsymbol{x} = \mathbf{A}^{-1}\boldsymbol{b} \\ \Rightarrow\ & \boldsymbol{x} = \mathbf{A}^{-1}\boldsymbol{b} . \end{aligned}$$

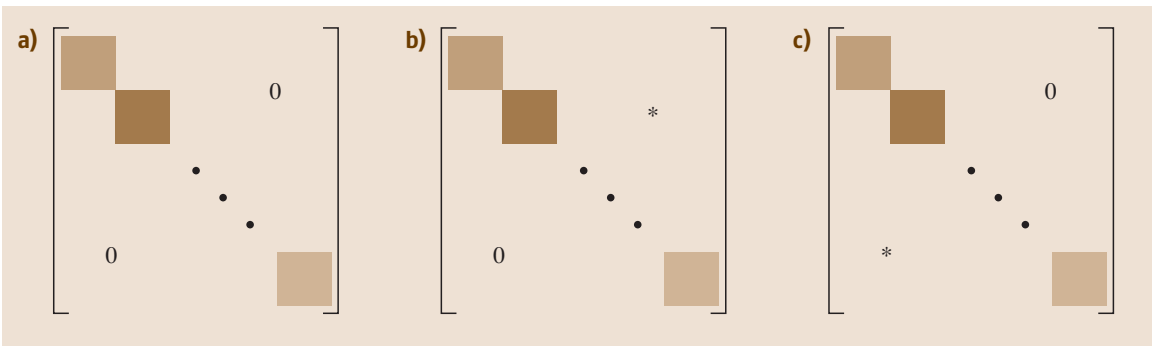


Fig. 1.23 (a) Block-diagonal matrix. (b) Block-upper-triangular matrix. (c) Block-lower-triangular matrix

Inverse via the Adjoint Matrix. The inverse of an invertible matrix $\mathbf{A} = [a_{ij}]_{n\times n}$ is determined using the adjoint of $\mathbf{A}$, denoted by adj($\mathbf{A}$) and defined as [1.1]

$$\begin{aligned}
&\text{adj}(\mathbf{A}) \\
&= \begin{pmatrix}
(-1)^{1+1}M_{11} & (-1)^{2+1}M_{21} & \cdots & (-1)^{n+1}M_{n1} \\
(-1)^{1+2}M_{12} & (-1)^{2+2}M_{22} & \cdots & (-1)^{n+2}M_{n2} \\
\vdots & \vdots & & \vdots \\
(-1)^{1+n}M_{1n} & (-1)^{2+n}M_{2n} & \cdots & (-1)^{n+n}M_{nn}
\end{pmatrix} \\
&= \begin{pmatrix}
C_{11} & C_{21} & \cdots & C_{n1} \\
C_{12} & C_{22} & \cdots & C_{n2} \\
\vdots & \vdots & & \vdots \\
C_{1n} & C_{2n} & \cdots & C_{nn}
\end{pmatrix} . \qquad (1.92)
\end{aligned}$$

Note that each minor M_{ij} (or cofactor C_{ij}) occupies the (j, i) position in the adjoint matrix, the opposite of what one would normally expect. Then, the inverse of $\mathbf{A}$ is simply defined by

$$\mathbf{A}^{-1} = \frac{1}{|\mathbf{A}|}\text{adj}(\mathbf{A}) . \qquad (1.93)$$

Example 1.33: Formula for the inverse of a 2×2 matrix
Find a formula for the inverse of

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} .$$

Solution. Following the procedure outlined above, we find

$$\begin{aligned}
M_{11} &= a_{22} , \quad & C_{11} &= a_{22} , \\
M_{12} &= a_{21} , \quad & C_{12} &= -a_{21} , \\
M_{21} &= a_{12} , \quad & C_{21} &= -a_{12} , \\
M_{22} &= a_{11} , \quad & C_{22} &= a_{11} .
\end{aligned}$$

Then,

$$\mathbf{A}^{-1} = \frac{1}{|\mathbf{A}|}\begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix} , \qquad (1.94)$$

which is a useful formula for 2×2 matrices, allowing us to omit the intermediate steps.

Inverses of Special Matrices. If the main diagonal entries are all nonzero, the inverse of a diagonal matrix is again diagonal. The diagonal elements of the inverse are simply the reciprocals of the diagonal elements of the original matrix. The inverse of an upper-triangular matrix is upper-triangular. The diagonal elements of the inverse are the reciprocals of the diagonal entries of the original matrix, while the off-diagonal entries do not obey any pattern. A similar result holds for lower-triangular matrices. Furthermore, it turns out that a block-diagonal matrix and its inverse have exactly the same structure.

Properties of Inverse. Some important properties of the inverse [1.1, 8] are given below. The assumption is that all listed inverses exist.

- $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$.
- $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$.
- $(\mathbf{A}^{\mathrm{T}})^{-1} = (\mathbf{A}^{-1})^{\mathrm{T}}$.
- The inverse of a symmetric matrix is symmetric.
- $(\mathbf{A}^{p})^{-1} = (\mathbf{A}^{-1})^{p}$, where p is a positive integer.
- $\det(\mathbf{A}^{-1}) = 1/\det(\mathbf{A})$.

1.5.2 Eigenvalues and Eigenvectors

The fundamentals of linear algebra are now extended to treat systems of differential equations, which are of particular importance to us since they represent the mathematical models of dynamic systems. In the analysis of such systems, one frequently encounters the *eigenvalue problem*, solutions of which are *eigenvalues* and *eigenvectors*. This knowledge enables the analyst to determine the natural frequencies and responses of systems. Let $\mathbf{A}$ be an $n\times n$ matrix, $\boldsymbol{v}$ a nonzero $n\times1$ vector, and λ a number (complex in general). Consider

$$\mathbf{A}\boldsymbol{v} = \lambda\boldsymbol{v} \qquad (1.95)$$

A number λ for which (1.95) has a nontrivial solution ($\boldsymbol{v} \neq \mathbf{0}_{n\times1}$) is called an eigenvalue or characteristic value of matrix $\mathbf{A}$. The corresponding solution $\boldsymbol{v} \neq \mathbf{0}$ of (1.95) is the eigenvector or characteristic vector of $\mathbf{A}$ corresponding to λ. Eigenvalues, together with eigenvectors form the eigensystem of $\mathbf{A}$. The problem of determining eigenvalues and the corresponding eigenvectors of $\mathbf{A}$, described by (1.95), is called an eigenvalue problem. The trace of a square matrix $\mathbf{A} = [a_{ij}]_{n\times n}$, denoted by tr($\mathbf{A}$), is defined as the sum of the eigenvalues of $\mathbf{A}$. It turns out that tr($\mathbf{A}$) is also the sum of the diagonal elements of $\mathbf{A}$. A matrix and its transpose have the same eigenvalues.

Solving the Eigenvalue Problem

Let us consider (1.95), $\mathbf{A}\boldsymbol{v} = \lambda\boldsymbol{v}$. Because equations in this form involve scalars, vectors, and matrices, it is im-

perative that extra caution is taken while working with them. First, rewrite and manipulate (1.95) as

$$\mathbf{A}\boldsymbol{v} - \lambda\boldsymbol{v} = \mathbf{0}_{n\times 1} \Rightarrow (\mathbf{A} - \lambda \boldsymbol{I})\boldsymbol{v} = \mathbf{0}\,, \tag{1.96}$$

where we note that every term here is an $n\times 1$ vector. The identity matrix $\mathbf{I} = \mathbf{I}_n$ has been inserted so that the two terms in parentheses are compatible; otherwise we would have $\mathbf{A} - \lambda$, which is meaningless. This equation has a nontrivial solution ($\boldsymbol{v} \neq \mathbf{0}$) if and only if the coefficient matrix, $\mathbf{A} - \lambda\mathbf{I}$, is singular. That means

$$|\mathbf{A} - \lambda\mathbf{I}| = 0\,. \tag{1.97}$$

This is called the characteristic equation of $\mathbf{A}$. The determinant $|\mathbf{A} - \lambda\mathbf{I}|$ is an nth-degree polynomial in λ and is known as the characteristic polynomial of $\mathbf{A}$ whose roots are precisely the eigenvalues of $\mathbf{A}$. Once the eigenvalues have been identified, each eigenvector corresponding to each of the eigenvalues is determined by solving (1.96).

Example 1.34: Eigenvalues and eigenvectors
Find the eigenvalues and eigenvectors of

$$\mathbf{A} = \begin{pmatrix} -1 & -3 \\ 0 & 2 \end{pmatrix}\,.$$

Solution. To find the eigenvalues of $\mathbf{A}$, we solve the characteristic equation,

$$\begin{aligned} &|\mathbf{A} - \lambda\mathbf{I}| = 0 \\ \Rightarrow &\begin{vmatrix} -1-\lambda & -3 \\ 0 & 2-\lambda \end{vmatrix} = 0 \\ &\Rightarrow (\lambda+1)(\lambda-2) = 0 \\ &\Rightarrow \lambda_{1,2} = -1, 2\,. \end{aligned}$$

Without losing any information, let us assign $\lambda_1 = -1$. To find the eigenvector, solve (1.96) with $\lambda = \lambda_1 = -1$,

$$(\mathbf{A} - \lambda_1\mathbf{I})\boldsymbol{v}_1 = \mathbf{0} \overset{\lambda_1 = -1}{\Rightarrow} (\boldsymbol{A} + \mathbf{I})\boldsymbol{v}_1 = \mathbf{0}\,, \tag{1.98}$$

where $\boldsymbol{v}_1$ is the 2×1 eigenvector corresponding to λ_1. Letting $\boldsymbol{v}_1 = \begin{pmatrix} v_{11} \\ v_{21} \end{pmatrix}$ and using $\mathbf{A}$ in (1.98), we find

$$\begin{pmatrix} 0 & -3 \\ 0 & 3 \end{pmatrix}\begin{pmatrix} v_{11} \\ v_{21} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}\,. \tag{1.99}$$

As expected, this system has nontrivial solutions because the coefficient matrix is singular. To solve (1.99), we apply suitable elementary row operations [1.1] to the augmented matrix to reduce it to

$$\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}\,.$$

The second row suggests that there is a free variable, implying that the two equations contained in (1.99) are linearly dependent. From the first row, we have $v_{21} = 0$ so that v_{21} cannot be the free variable, so v_{11} must be. In this example, since we already have $v_{21} = 0$, then $v_{11} \neq 0$ because otherwise $\boldsymbol{v}_1 = \mathbf{0}$, which is not valid. For simplicity, let $v_{11} = 1$, so

$$\boldsymbol{v}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}\,.$$

Similarly, the eigenvector corresponding to $\lambda_2 = 2$ can be shown to be $\boldsymbol{v}_2 = [-1\ 1]^{\mathrm{T}}$. The set $(\boldsymbol{v}_1, \boldsymbol{v}_2)$ is the basis of all eigenvectors of matrix $\mathbf{A}$.

Special Matrices

The eigenvalues of triangular and diagonal matrices are the diagonal entries. The eigenvalues of block-triangular and diagonal matrices are the eigenvalues of the block matrices along the main diagonal. All eigenvalues of a symmetric matrix are real, while those of a skew-symmetric matrix are either zero or pure imaginary.

Generalized Eigenvectors

If λ_k is an eigenvalue of $\mathbf{A}$ occurring m_k times, then m_k is the algebraic multiplicity of λ_k, denoted by $\mathrm{AM}(\lambda_k)$. The maximum number of linearly independent eigenvectors associated with λ_k is called the geometric multiplicity of λ_k, $\mathrm{GM}(\lambda_k)$. In general, $\mathrm{GM}(\lambda_k) \leq \mathrm{AM}(\lambda_k)$. In Example 1.34 the AM and GM of each of the two eigenvalues was 1. When $\mathrm{GM}(\lambda_k) < \mathrm{AM}(\lambda_k)$, there are fewer eigenvectors than one would expect. For instance, if $\mathrm{AM}(\lambda) = 2$ and $\mathrm{GM}(\lambda) = 1$, then only one independent eigenvector can be found for λ, while one is missing; the missing one is called a generalized eigenvector [1.1].

Similarity Transformation – Diagonalization

Two matrices $\mathbf{A}_{n\times n}$ and $\mathbf{B}_{n\times n}$ are said to be similar if there exists a nonsingular matrix $\boldsymbol{S}_{n\times n}$ such that $\mathbf{B} = \mathbf{S}^{-1}\mathbf{A}\mathbf{S}$. We say that $\mathbf{B}$ is obtained from $\mathbf{A}$ through a similarity transformation. Similar matrices have the same eigenvalues. Suppose $\mathbf{A}_{n\times n}$ has n linearly independent eigenvectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_n$ associated with eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_n$. Form the $n\times n$ matrix $\mathbf{P} = [\ \boldsymbol{v}_1\ \ \boldsymbol{v}_2\ \ .\ .\ \boldsymbol{v}_n\ ]$, known as the modal matrix, whose columns are the eigenvectors of $\mathbf{A}$. Then, $\mathbf{P}$

is nonsingular, and

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} , \qquad (1.100)$$

where the order of the appearance of the λ_i along the main diagonal agrees with the order of the corresponding $\boldsymbol{v}_i$ in matrix $\mathbf{P}$. Any matrix $\mathbf{A}$ that satisfies (1.100) is called diagonalizable. It is clear that $\mathbf{A}$ and $\mathbf{\Lambda}$ are similar, hence share the same eigenvalues.

Example 1.35: Modal matrix
In reference to Example 1.34, the modal matrix is

$$\mathbf{P} = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} , \text{ which satisfies}$$

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{\Lambda} = \begin{pmatrix} -1 & 0 \\ 0 & 2 \end{pmatrix} = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} .$$

1.5.3 Numerical Solution of Higher-Order Systems of ODEs

In this section we present a two-step process to treat one or more ODEs of order higher than one. The first stage is concerned with the conversion of the original ODEs into a system of first-order ODEs. The second stage deals with the application of the fourth-order Runge–Kutta method to solve this system numerically; also see Sect. 1.2.2. It should be mentioned that we are not interested in mathematically fabricated, higher-order systems, but rather those of physical significance. The first step is handled via the state variables, explained below.

State Variables

The smallest possible set of independent variables that completely describes the state of a system is referred to as the set of state variables. Since independence is essential, state variables cannot be expressible as algebraic functions of one another. Moreover, the set of state variables for a certain system is *not unique*. Given a set of differential equations describing a certain physical system, two key questions need be answered [1.1, 3].

1. How many state variables are there?
   The number of state variables is equal to the number of initial conditions required to completely solve the system's governing equations.
2. What are chosen as the state variables?
   Those variables for which initial conditions are required in (1) are chosen as the state variables.

Example 1.36: State variables
The equation of motion for the mass–spring–damper system in Fig. 1.24 is given as

$$m\ddot{x} + c\dot{x} + kx = f(t) .$$

For a complete description, we need the initial conditions $x(0) = x_0$ (initial displacement) and $\dot{x}(0) = v_0$ (initial velocity).

Since two initial conditions are required, there are two state variables. And, since these conditions are required for x and $\dot{x}$, the state variables are

$$x_1 = x , \quad x_2 = \dot{x} .$$

State-Variable Equations. The idea now is to use the selected set of state variables to transform the original system into a larger system of first-order ODEs. Each of the resulting differential equations consists of the time derivative of one of the state variables on one side, and a linear combination of the state variables and possibly system inputs on the other side. Each of these first-order ODEs just described is called a state-variable equation.

Example 1.37: State-variable equations
Referring to Example 1.36, since there are two state variables, two first-order ODEs must be obtained in terms of x_1 and x_2. The first equation is simply $\dot{x}_1 = x_2$. To obtain the second equation, we use the equation of motion as

$$\dot{x}_2 = \frac{1}{m}[-kx_1 - cx_2 + f(t)] .$$

The state-variable equations are then

$$\begin{cases} \dot{x}_1 = x_2 \\ \dot{x}_2 = \frac{1}{m}[-kx_1 - cx_2 + f(t)] . \end{cases}$$

These can be expressed in matrix form, as

$$\dot{\boldsymbol{x}} = \boldsymbol{f}(t, \boldsymbol{x}) , \boldsymbol{x}(0) = \boldsymbol{x}_0 , \qquad (1.101)$$

where

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} ,$$

$$\boldsymbol{f}(t, \boldsymbol{x}) = \begin{pmatrix} x_2 \\ -(k/m)x_1 - (c/m)x_2 + (1/m)f(t) \end{pmatrix} ,$$

$$\boldsymbol{x}_0 = \begin{pmatrix} x_0 \\ v_0 \end{pmatrix} .$$

Fourth-Order Runge–Kutta Method for Systems

Numerical solution of the state-variable equations – such as that in (1.101) of Example 1.37 – is then obtained via the extension of RK4 discussed in Sect. 1.2.2. Consider a system in the form

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{f}(t, \boldsymbol{x}(t)) \,, \quad \boldsymbol{x}(a) = \boldsymbol{x}_0 \,, \quad a \le t \le b \,, \tag{1.102}$$

where

$$\boldsymbol{x}(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_n(t) \end{pmatrix} \,,$$

$$\boldsymbol{f}(t, \boldsymbol{x}(t)) = \begin{pmatrix} f_1(t, x_1, x_2, \cdots, x_n) \\ f_2(t, x_1, x_2, \cdots, x_n) \\ \vdots \\ f_n(t, x_1, x_2, \cdots, x_n) \end{pmatrix} \,,$$

$$\boldsymbol{x}_0 = \boldsymbol{x}(a) = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{pmatrix} \,.$$

Define an integer $N > 0$ and let $h = (b-a)/N$ be the step size. The mesh points $t_i = a + ih$, $i = 0, 1, \cdots, N-1$, then partition the interval $[a, b]$ into

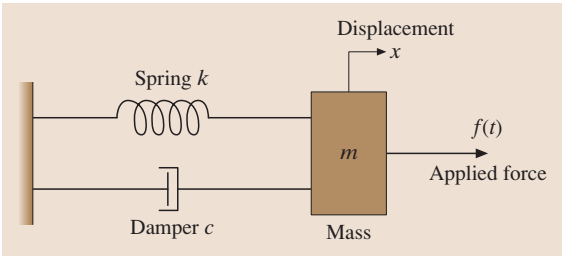


Fig. 1.24 A mechanical system

N subintervals. The fourth-order Runge–Kutta method (RK4) for a system of first-order ODEs is as follows [1.5]. Knowing the initial vector $\boldsymbol{x}_0$, the solution vector $\boldsymbol{x}_i$ at each of the subsequent mesh points t_i is obtained via

$$\boldsymbol{x}_{i+1} = \boldsymbol{x}_i + \frac{1}{6}[\boldsymbol{q}_1 + 2\boldsymbol{q}_2 + 2\boldsymbol{q}_3 + \boldsymbol{q}_4], \quad i = 0, 1, 2, \cdots, N-1 \,,$$

where

$$\boldsymbol{q}_1 = h\boldsymbol{f}(t_i, \boldsymbol{x}_i) \,,$$

$$\boldsymbol{q}_2 = h\boldsymbol{f}\left(t_i + \frac{1}{2}h, \boldsymbol{x}_i + \frac{1}{2}\boldsymbol{q}_1\right) \,,$$

$$\boldsymbol{q}_3 = h\boldsymbol{f}\left(t_i + \frac{1}{2}h, \boldsymbol{x}_i + \frac{1}{2}\boldsymbol{q}_2\right) \,,$$

$$\boldsymbol{q}_4 = h\boldsymbol{f}(t_i + h, \boldsymbol{x}_i + \boldsymbol{q}_3) \,.$$

References

1.1 R.S. Esfandiari: *Applied Mathematics for Engineers*, 4th edn. (Atlantis, Irvine, California 2008)

1.2 J.W. Brown, R.V. Churchill: *Complex Variables and Applications*, 7th edn. (McGraw-Hill, New York 2003)

1.3 H.V. Vu, R.S. Esfandiari: *Dynamic Systems: Modeling and Analysis* (McGraw-Hill, New York 1997)

1.4 C.H. Edwards, D.E. Penney: *Elementary Differential Equations with Boundary Value Problems*, 4th edn. (Prentice-Hall, New York 2000)

1.5 J.H. Mathews: *Numerical Methods for Computer Science, Engineering, and Mathematics* (Prentice-Hall, New York 1987)

1.6 R.L. Burden, J.D. Faires: *Numerical Analysis*, 3rd edn. (Prindle, Boston 1985)

1.7 J.W. Brown, R.V. Churchill: Fourier Series and Boundary Value Problems. 6th edn. (McGraw-Hill, New York 2001)

1.8 G.H. Golub, C.F. Van Loan: *Matrix Computations*, 3rd edn. (The Johns Hopkins University Press, London 1996)

2. Mechanics

Hen-Geul Yeh, Hsien-Yang Yeh, Shouwen Yu

Mechanics is the study of the motion of matter and the forces that cause such motion, and is based on the concepts of time, space, force, energy, and matter. A knowledge of mechanics is needed for the study of all branches of physics, chemistry, biology, and engineering [2.1]. The subject of mechanics is logically divided into two parts: statics, which is concerned with the equilibrium of bodies under the action of forces, and dynamics, which is concerned with the motion of bodies. The principles of mechanics as a science are rigorously expressed by mathematics, which therefore plays an important role in the application of these principles to the solution of practical problems [2.2]. A force is a vector quantity, because its effect depends on the direction as well as on the magnitude of the action. In addition to the tendency to move a body in the direction of its application, a force can also tend to rotate a body about an axis. This rotational tendency is known as the moment of the force and therefore, moment can be expressed as a vector quantity as well. When a body is in equilibrium, the resultant of all forces acting on it is zero. Thus, the resultant force and the resultant moment are both zero and the equilibrium equations are satisfied.

A large number of problems involving actual structures, however, can be reduced to problems concerning the equilibrium of a particle. This is done by choosing a significant particle and drawing a separate diagram showing this particle and all the forces acting on it. Such a diagram is called a free-body diagram. The same concept is applied to the solution of a rigid-body equilibrium problem as well [2.3]. A truss is a structure composed of (usually straight) members joined together at their end points and loaded only at the joints. Trusses are commonly seen supporting the roofs of buildings as well as large railroad and highway bridges [2.4]. The analysis of truss structures is a typical engineering application of statics. To analyze systems of forces distributed over an area or volume, we have to evaluate the centroids and center of gravity as well as moments of inertia.

Consider a practical question: what is the steepest incline on which a truck can be parked without slipping? To answer this question, we must examine the nature of friction forces in more detail. Eventually the first variational principle we encounter in the science of mechanics is that of virtual work, which controls the equilibrium of a mechanical system and is fundamental to the development of analytical mechanics.

Dynamic mechanics can be divided into two parts: (1) kinematics, which is the study of a geometry of motion and is used to relate displacement, velocity, acceleration, and time, without taking into account forces and moments as causes of the motion, and (2) dynamics, which is the study of the relation between the forces and moments acting on a body, and the mass and motion of the body; it is used to predict the motion caused by given forces and moments or to determine the forces and moments required to produce a given motion.

This chapter is also devoted to kinematics, which is the starting point from which begin the analysis of the basic motion of particles and rigid bodies and the dynamics of a single particle. This is a fundamental concept in which Newton's laws and certain principles of dynamics are introduced. Furthermore, advanced materials, such as the dynamics of systems of particles, momentum equations, Lagrange's equations, energy equations, D'Alembert's principle, and the dynamics of rigid bodies are also included. Lagrange's equations of motion for linear systems are introduced at the end of the chapter, although this can be regarded as the beginning of the vibration.

2.1 Statics of Rigid Bodies

Statics is a branch of classical mechanics, which is part of the foundation of physics and modern engineering technology. Statics is the study of the equilibrium of rigid bodies under the action of forces and moments. According to Newton's Laws, equilibrium prevails if a body is at rest or is in uniform motion along a straight line.

A rigid body can be represented as a collection of particles. The size and shape of a rigid body remain constant at all times and under all loading conditions. In other words, rigid bodies as understood in statics as bodies of which the deformations are so small that the points at which force is applied undergo negligible displacement.

2.1.1 Force

A force represents the action of one body on another and is generally characterized by its point of application, magnitude, and direction. Thus, force is a vector quantity.

Introducing the unit vectors $\boldsymbol{e}_x$, $\boldsymbol{e}_y$, and $\boldsymbol{e}_z$, or $\boldsymbol{i}$, $\boldsymbol{j}$, and $\boldsymbol{k}$, directed along the x, y, and z axes, respectively, the force $\boldsymbol{F}$ can be expressed in the form (Fig. 2.1a,b)

$$\begin{aligned}\boldsymbol{F} &= F_x\boldsymbol{e}_x + F_y\boldsymbol{e}_y + F_z\boldsymbol{e}_z \\ &= F_x\boldsymbol{i} + F_y\boldsymbol{j} + F_z\boldsymbol{k} \\ &= (F\cos\alpha)\,\boldsymbol{e}_x + (F\cos\beta)\,\boldsymbol{e}_y + (F\cos\gamma)\,\boldsymbol{e}_z \\ &= (F\cos\theta_x)\,\boldsymbol{i} + \left(F\cos\theta_y\right)\boldsymbol{j} + (F\cos\theta_z)\,\boldsymbol{k}, \end{aligned} \quad (2.1)$$

where

$$F = |F| = \sqrt{F_x^2 + F_y^2 + F_z^2}\,. \quad (2.2)$$

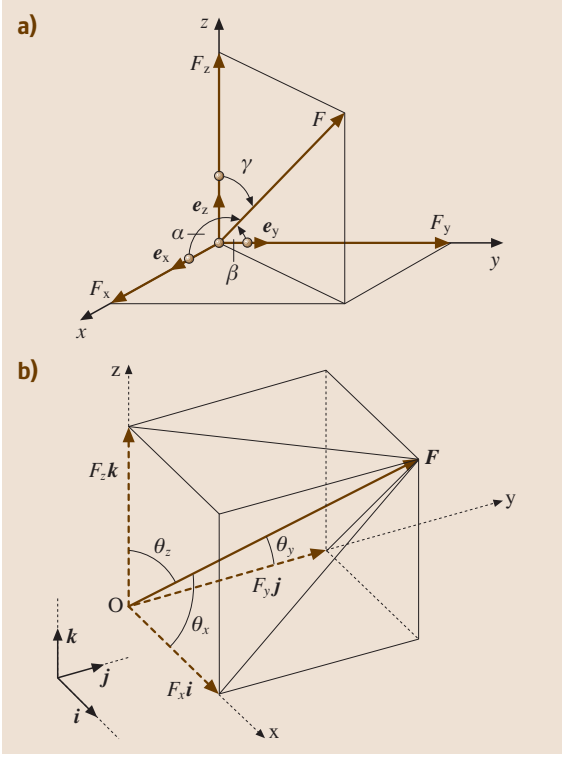


Fig. 2.1a,b Vector representation of a force F

The direction cosines are defined as $\cos\alpha = \cos\theta_x = F_x/F$, $\cos\beta = \cos\theta_y = F_y/F$, $\cos\gamma = \cos\theta_z = F_z/F$, and $\cos^2\alpha + \cos^2\beta + \cos^2\gamma = \cos^2\theta_x + \cos^2\theta_y + \cos^2\theta_z = 1$. When solving three-dimensional problems, it is necessary to find the x, y, and z scalar components of a force. In most cases, the direction of a force is described by two points on the line of action of the force (Fig. 2.1a), or by two angles which orient the line of action (Fig. 2.1b):

1. The direction of a force $\boldsymbol{F}$ is defined by the coordinates of two points $M(x_1, y_1, z_1)$ and $N(x_2, y_2, z_2)$, located on its line of action (Fig. 2.2a). Consider the vector $\boldsymbol{MN}$ joining M and N and of the same sense as $\boldsymbol{F}$. Therefore

$$\boldsymbol{MN} = (x_2 - x_1)\boldsymbol{i} + (y_2 - y_1)\boldsymbol{j} + (z_2 - z_1)\boldsymbol{k} \,. \tag{2.3}$$

The unit vector $\boldsymbol{\lambda}$ along the line of action of $\boldsymbol{F}$ is obtained by dividing the vector $\boldsymbol{MN}$ by its magnitude MN, thus

$$\begin{aligned}\boldsymbol{\lambda} &= \frac{\boldsymbol{MN}}{\mathrm{MN}} \\ &= \frac{(x_2 - x_1)\boldsymbol{i} + (y_2 - y_1)\boldsymbol{j} + (z_2 - z_1)\boldsymbol{k}}{\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}} \\ &= \frac{d_x\boldsymbol{i} + d_y\boldsymbol{j} + d_z\boldsymbol{k}}{d} \,,\end{aligned} \tag{2.4}$$

and $\boldsymbol{F} = F\boldsymbol{\lambda}$, where F is the magnitude of the force $\boldsymbol{F}$,

$$\begin{aligned} &d_x = x_2 - x_1 \,, \quad d_y = y_2 - y_1 \,, \\ &d_z = z_2 - z_1 \quad \text{and} \\ &d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2} \,. \end{aligned}$$

The angles θ_x, θ_y, and θ_z that $\boldsymbol{F}$ forms with the coordinate axes can be expressed as

$$\cos\theta_x = \frac{d_x}{d} \,, \quad \cos\theta_y = \frac{d_y}{d} \,, \quad \cos\theta_z = \frac{d_z}{d} \,.$$

2. Consider the geometry of Fig. 2.2b, assuming that the angles θ and ϕ are known. First resolve $\boldsymbol{F}$ into its horizontal and vertical components:

$$F_{xy} = F\cos\phi \,, \quad F_z = F\sin\phi \,.$$

Then resolve the horizontal components F_{xy} into the x-and y-components

$$\begin{aligned} F_x &= F_{xy}\cos\theta = F\cos\phi\cos\theta \,, \\ F_y &= F_{xy}\sin\theta = F\cos\phi\sin\theta \,. \end{aligned}$$

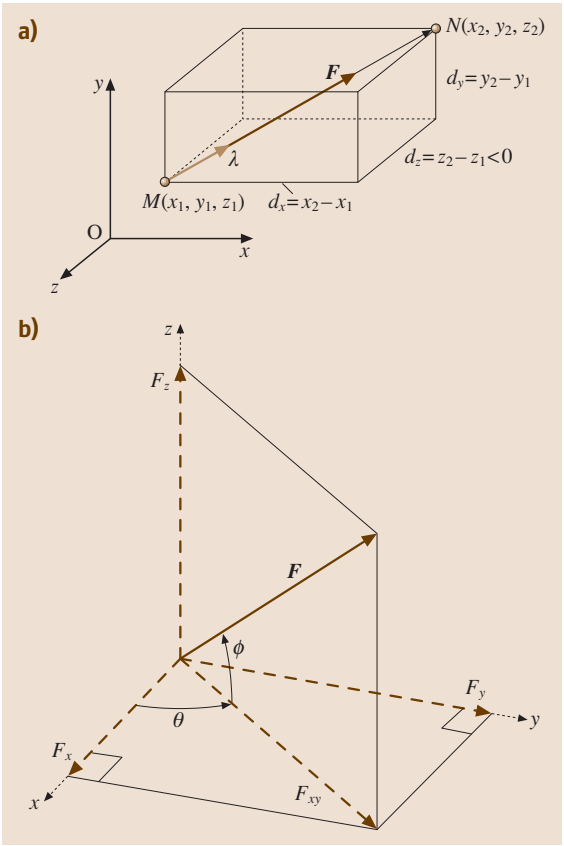


Fig. 2.2a,b Force defined by its magnitude and two points on its line of action

The quantities F_x, F_y, and F_z are the desired scalar components of $\boldsymbol{F}$. Forces acting on rigid bodies can be separated into two groups: (a) external, and (b) internal forces. The external forces represent the action of other bodies on the rigid body under consideration. The internal forces are the forces that hold together the particles forming the rigid body.

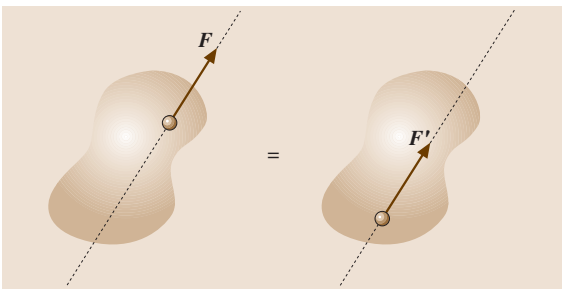


Fig. 2.3 Principle of transmissibility

The principle of transmissibility states that the conditions of equilibrium or motion of a rigid body remain unchanged if a force $\boldsymbol{F}$ acting at a given point on the rigid body is replaced by a force $\boldsymbol{F}'$ of the same magnitude and same direction, but acting at a different point, provided that the two forces have the same line of action (Fig. 2.3); these are known as equivalent forces.

2.1.2 Addition of Concurrent Forces in Space and Equilibrium of a Particle

The resultant $\boldsymbol{R}$ of two or more forces in space is usually determined by summing their rectangular components. Graphical or trigonometric methods are generally not practical in the case of forces in space

$$\boldsymbol{R} = \sum \boldsymbol{F} = 0\,, \tag{2.5}$$

$$\begin{aligned} R_x\boldsymbol{e}_x + R_y\boldsymbol{e}_y + R_z\boldsymbol{e}_z &= R_x\boldsymbol{i} + R_y\boldsymbol{j} + R_z\boldsymbol{k} \\ &= \left(\sum F_x\right)\boldsymbol{i} + \left(\sum F_y\right)\boldsymbol{j} \\ &\quad + \left(\sum F_z\right)\boldsymbol{k} \\ &= \left(\sum F_x\right)\boldsymbol{e}_x + \left(\sum F_y\right)\boldsymbol{e}_y \\ &\quad + \left(\sum F_z\right)\boldsymbol{e}_z\,. \end{aligned} \tag{2.6}$$

From which it follows that

$$R_x = \sum F_x\,, \quad R_y = \sum F_y\,, \quad R_z = \sum F_z\,. \tag{2.7}$$

The magnitude of the resultant and the angles θ_x, θ_y, and θ_z that the resultant forms with the coordinate axes are

$$R = \sqrt{R_x^2 + R_y^2 + R_z^2}\,, \tag{2.8}$$

$$\cos\theta_x = \frac{R_x}{R}\,, \quad \cos\theta_y = \frac{R_y}{R}\,, \quad \cos\theta_z = \frac{R_z}{R}\,. \tag{2.9}$$

Statics deals primarily with the description of the force conditions necessary and sufficient to maintain the equilibrium of engineering structures. When a body is in equilibrium, the resultant of all the forces acting on it is zero. Therefore, for the equilibrium of a particle in space

$$\sum F_x = 0\,, \quad \sum F_y = 0\,, \quad \sum F_z = 0\,. \tag{2.10}$$

2.1.3 Moment and Couple

In addition to the tendency to move a body in the direction of its application, a force can also tend to rotate a body about an axis. The axis may be any line that neither intersects nor is parallel to the line of action of the force. This rotational tendency is known as the moment M of the force (Fig. 2.4a).

The moment produced by two equal, opposite, and non-collinear forces is called a couple (Fig. 2.4b).

Consider a force $\boldsymbol{F}$ acting on a rigid body as shown in Fig. 2.4a. The position vector $\boldsymbol{r}$ and the force $\boldsymbol{F}$ define the plane A. The moment $\boldsymbol{M}_O$ of $\boldsymbol{F}$ about an axis through O normal to the plane has magnitude $\boldsymbol{M}_O = Fd$, where d is the perpendicular distance from O to the line of $\boldsymbol{F}$. This

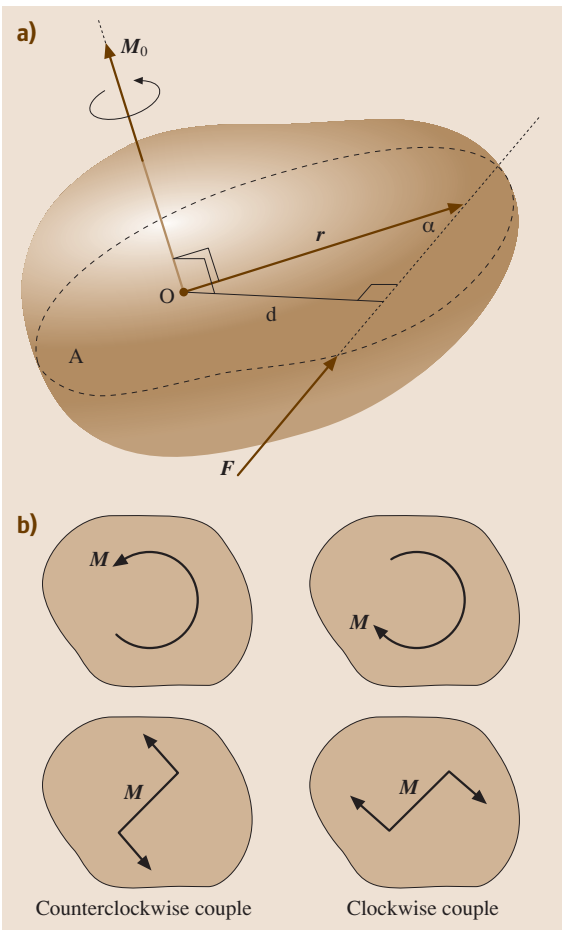


Fig. 2.4 (a) Moment of a force about a point, (b) couple

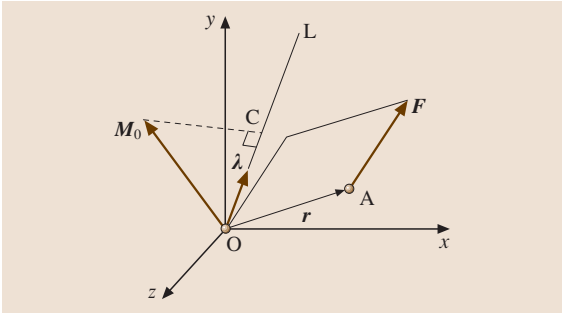


Fig. 2.5 Moment about an arbitrary axis

moment is referred to as the moment of $\boldsymbol{F}$ about the point O.

The vector form of the moment of $\boldsymbol{F}$ about point O is

$$\boldsymbol{M}_O = \boldsymbol{r}\times\boldsymbol{F} = \begin{vmatrix} \boldsymbol{i} & \boldsymbol{j} & \boldsymbol{k} \\ r_x & r_y & r_z \\ F_x & F_y & F_z \end{vmatrix} . \qquad (2.11a)$$

Expansion of the determinant gives

$$\boldsymbol{M}_O = (r_y F_z - r_z F_y)\boldsymbol{i} + (r_z F_x - r_x F_z)\boldsymbol{j} + (r_x F_y - r_y F_x)\boldsymbol{k} \qquad (2.11b)$$

The moment about an arbitrary axis can be found by considering a force $\boldsymbol{F}$ acting on a rigid body and the moment $\boldsymbol{M}_O$ of that force about O, as shown in Fig. 2.5. Let OL be an axis through O, then the moment $\boldsymbol{M}_{OL}$ of $\boldsymbol{F}$ about OL as the projection OC of the moment $\boldsymbol{M}_O$ onto the axes OL is defined as

$$\boldsymbol{M}_{OL} = \boldsymbol{\lambda}\cdot\boldsymbol{M}_O = \boldsymbol{\lambda}\cdot(\boldsymbol{r}\times\boldsymbol{F}) , \qquad (2.12a)$$

which shows that the moment $\boldsymbol{M}_{OL}$ of $\boldsymbol{F}$ about the axis OL is the scalar obtained by the triple scalar product of $\boldsymbol{\lambda}, \boldsymbol{r}$, and $\boldsymbol{F}$.

Expressing $\boldsymbol{M}_{OL}$ in the form of a determinant

$$\boldsymbol{M}_{OL} = \begin{vmatrix} \lambda_x & \lambda_y & \lambda_z \\ x & y & z \\ F_x & F_y & F_z \end{vmatrix} , \qquad (2.12b)$$

where λ_x, λ_y, and λ_z are the direction cosines of the axis OL. x, y, and z are the coordinates of the point of application of $\boldsymbol{F}$, and F_x, F_y, and F_z are the components of the force $\boldsymbol{F}$.

2.1.4 Equilibrium Conditions

When a body is in equilibrium, the resultant of all forces acting on it is zero. Thus, the resultant force $\boldsymbol{R}$ and the resultant moment $\boldsymbol{M}$ are both zero, and the equilibrium equations result

$$\boldsymbol{R} = \sum \boldsymbol{F} = 0 , \quad \boldsymbol{M} = \sum \boldsymbol{M} = \sum (\boldsymbol{r}\times\boldsymbol{F}) = 0 . \qquad (2.13)$$

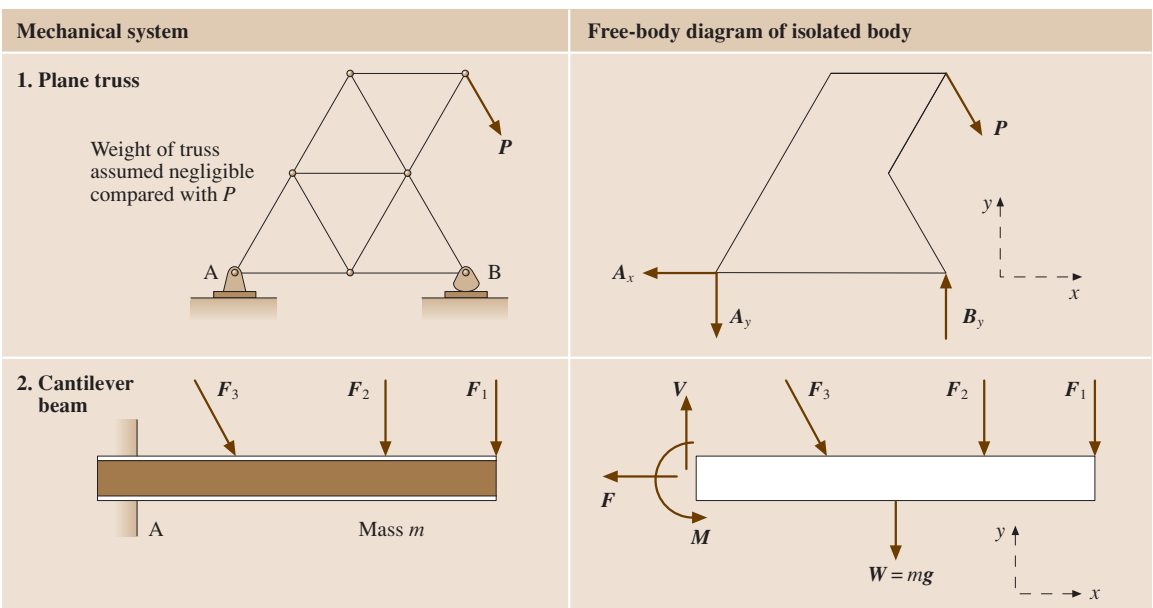


Fig. 2.6 Free-body diagrams (after [2.2])

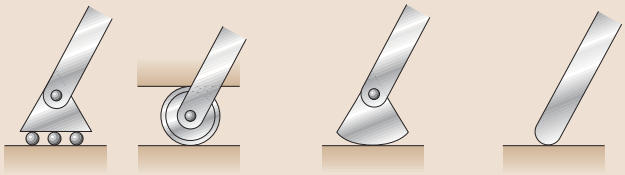
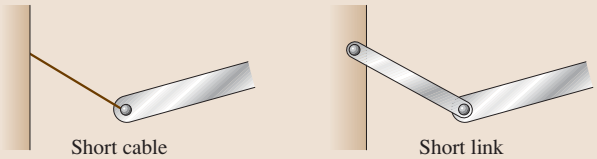
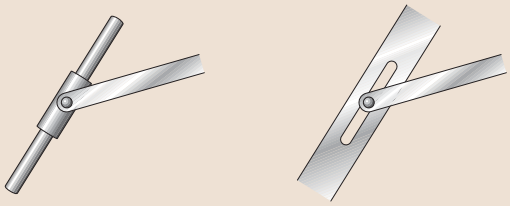
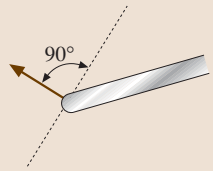
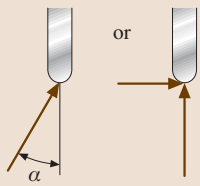
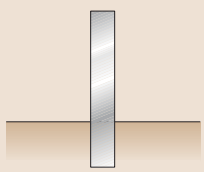
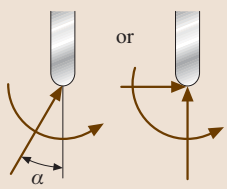
Support or connection	Reaction	Number of unknowns
 Rollers; Rocker; Frictionless surface	 Force with known line of action	1
 Short cable; Short link	 Force with known line of action	1
 Collar on frictionless rod; Frictionless pin in slot	 Force with known line of action	1
 Frictionless pin or hinge; Rough surface	 Force of unknown direction	2
 Fixed support	 Force and couple	3

Fig. 2.7 Reactions at supports and connections for a two-dimensional structure (after [2.3])

These requirements are both necessary and sufficient conditions for equilibrium.

In statics, the primary concern is to study forces that act on rigid bodies at rest. In solving a problem concerning the equilibrium of a rigid body, it is essential to consider all of the forces acting on the body. Therefore, the first step in the solution of the problem should be to draw a free-body diagram of the rigid body under consideration. The free-body diagram is a diagrammatic representation of the isolated system treated as a single body. The diagram shows all forces applied to the system by mechanical contact with other bodies, which are imagined to be removed. Examples of free-body diagrams, and reactions at supports are shown in Figs. 2.6–2.8.

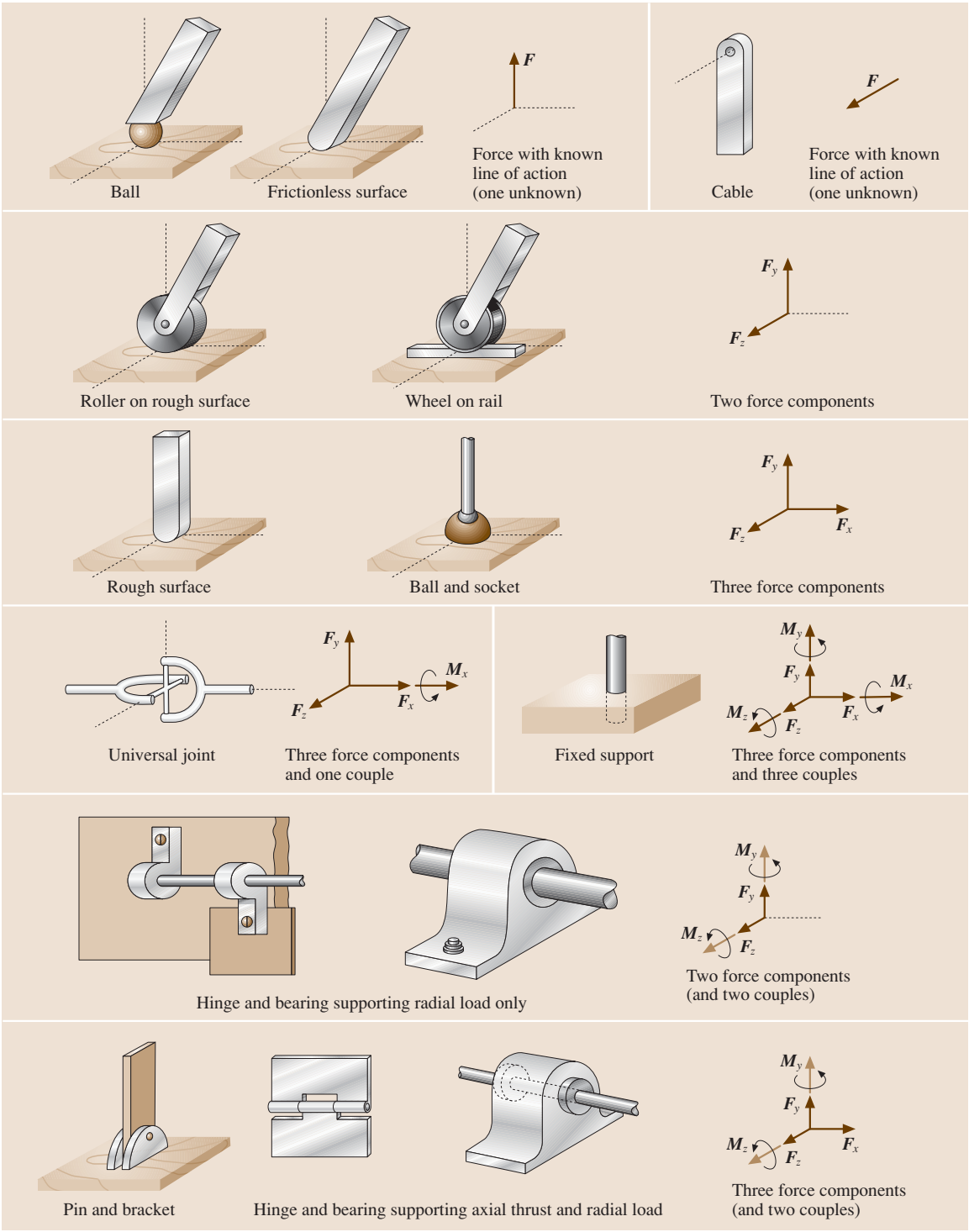


Fig. 2.8 Reactions at supports and connections for a three-dimensional structure (after [2.3])

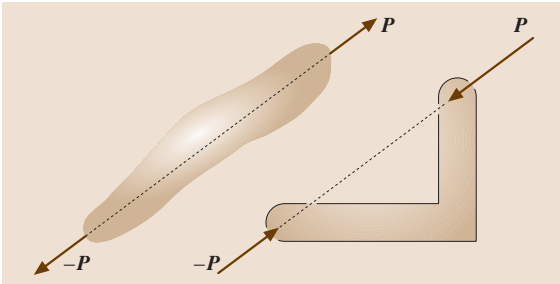


Fig. 2.9 Two-force member

There are two frequently occurring equilibrium situations: two- and three-force members. A two-force member is the equilibrium of a body under the action of two forces only, as shown in Fig. 2.9. For a two-force member to be in equilibrium, the forces must be equal, opposite and collinear. The shape of the member does not affect this simple requirement.

A three-force member is a body under the action of three forces, as shown in Fig. 2.10. For a three-force member to be in equilibrium, the lines of action of the three forces must be concurrent. The principle of the concurrency of three forces in equilibrium is of considerable use in carrying out a graphical solution of the force equations. In this case, the polygon of forces is drawn and made to close, as shown in Fig. 2.8b.

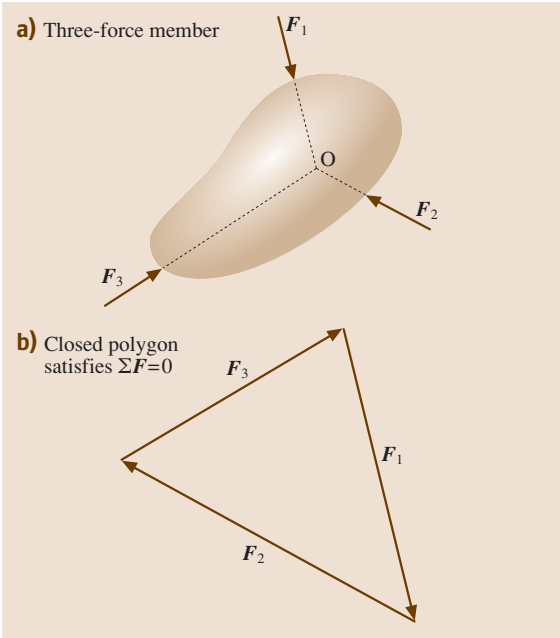


Fig. 2.10a,b Three-force member

A rigid body that possesses more external supports or constraints than are necessary to maintain an equilibrium position is called statistically indeterminate. Supports that can be removed without destroying the equilibrium condition of the body are said to be redundant. The number of redundant supporting elements corresponds to the degree of statical indeterminacy and equals the total number of unknown external forces minus the number of available independent equations of equilibrium. On the other hand, bodies supported by the minimum number of constraints necessary to ensure an equilibrium configuration are called statically determinate, and for such bodies the equilibrium equations are sufficient to determine the unknown external forces.

2.1.5 Truss Structures

A framework composed of members joined at their ends to form a rigid structure is called a truss. When the members of the truss lie essentially in a single plane, the truss is called a plane truss. The basic element of a plane truss is the triangle. Three bars joined by pins at their ends constitute a rigid truss and a larger rigid truss can be obtained by adding two new members to the first one and connecting them at a new joint as shown in Fig. 2.11. Trusses obtained by repeating this procedure are called simple trusses. One may check that, in a simple truss, the total number of members m is expected by $m = 2n - 3$, where n is the total number of joints. This expression applies to a statically determinate and stable truss, since two conditions of equilibrium exist for each joint; i. e., of the $2n - 3$ conditions of equilibrium, m unknown axial forces can be calculated. A pin-jointed truss with $m < 2n - 3$ members is statically undeterminate and kinematically unstable, and a pin-jointed truss with $m > 2n - 3$ is internally statically indeterminate.

The forces in the various members of a simple truss can be determined by the method of joints. First the reactions at the supports can be obtained by considering the entire truss as a free body. The free-body diagram of each pin is then drawn, showing the forces exerted on the pin by the members or supports it connects. Since the members are straight two-force members, the force exerted by a member on the pin is directed along that member and only the magnitude of the force is unknown. It is always possible in the case of a simple truss to draw the free-body diagrams of the pins in such an order that only two unknown forces are included in each diagram. These forces can be determined from the

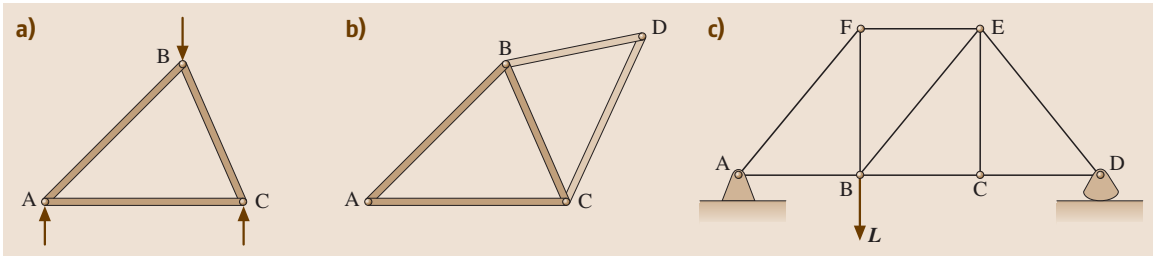


Fig. 2.11a–c Truss structure

two corresponding equilibrium equations: $\sum F_x = 0$ and $\sum F_y = 0$. If the force exerted by a member on a pin is directed toward that pin, the member is in compression; if it is directed away from the pin, the member is in tension.

The method of section is usually preferred to the method of joints when the force in only one member or very few members of a truss is desired. For example, to determine the force in member BD of the truss shown in Fig. 2.12, it is better to pass a section through members BD, BE, and CE, remove these members and use the portion ABC of the truss as a free body. Writing $\sum M_E = 0$ to determine the magnitude of the force F_{BD}, which represents the force in member BD. A positive sign indicates that the member is in tension and a negative sign indicates that it is in compression.

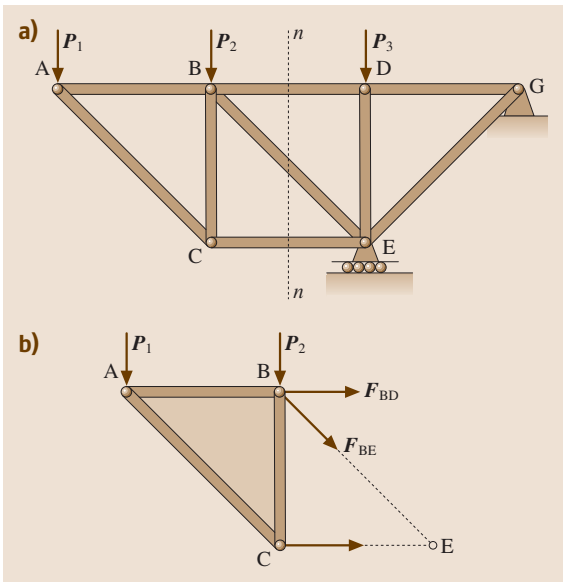


Fig. 2.12a,b Method of section

2.1.6 Distributed Forces

The mass elements of a body of mass m are affected by the forces of gravity $dF = dmg = dW$, all of which are parallel to one another. To determine the location of the center of gravity of any body mathematically, we note that the moment of the resultant gravitational force W about any axis equals the sum of the moments of the gravitational forces dW acting on all the particles treated as infinitesimal elements of the body about the same axis. The resultant of the gravitational forces acting on all elements is the weight of the body and is given by the sum $W = \int dW$. For example, as shown in Fig. 2.13, the moment about the y-axis of the elemental weight is $x\,dW$, and the sum of these moments for all elements of the body is $\int x\,dW$. This sum of moments must equal $W\overline{x}$, the moment of the sum.

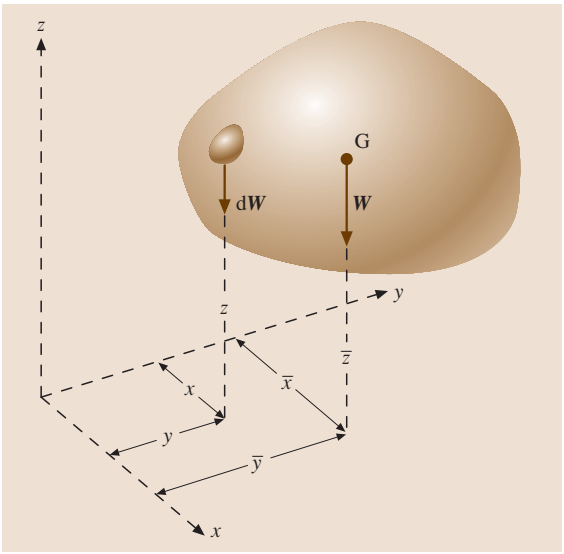


Fig. 2.13 Center of gravity

Therefore, $\overline{x}W = \int x\,\mathrm{d}W$. With similar expressions for the other two components, the coordinates of the center of gravity G are expressed as

$$\overline{x} = \frac{\int x\,\mathrm{d}W}{W}\,, \quad \overline{y} = \frac{\int y\,\mathrm{d}W}{W}\,, \quad \overline{z} = \frac{\int z\,\mathrm{d}W}{W}\,. \tag{2.14}$$

With the substitutions $W = mg$ and $\mathrm{d}W = g\,\mathrm{d}m$, the expressions for the coordinates of the center of gravity become

$$\overline{x} = \frac{\int x\,\mathrm{d}m}{m}\,, \quad \overline{y} = \frac{\int y\,\mathrm{d}m}{m}\,, \quad \overline{z} = \frac{\int z\,\mathrm{d}m}{m}\,. \tag{2.15}$$

The density ρ of a body is its mass per unit volume. Therefore, the mass of a differential element of volume $\mathrm{d}V$ becomes $\mathrm{d}m = \rho\,\mathrm{d}V$, thus, (2.15) can be written as

$$\overline{x} = \frac{\int x\rho\,\mathrm{d}V}{\int \rho\,\mathrm{d}V}\,, \quad \overline{y} = \frac{\int y\rho\,\mathrm{d}V}{\int \rho\,\mathrm{d}V}\,, \quad \overline{z} = \frac{\int z\rho\,\mathrm{d}V}{\int \rho\,\mathrm{d}V}\,. \tag{2.16}$$

Since g no longer appears in (2.15) and (2.16), a unique point that is only a function of the distribution of mass is defined in the body. This point is called the center of mass and clearly coincides with the center of gravity as long as the gravity field is treated as uniform and parallel.

The calculation of centroids falls within three distinct categories:

1. Lines: for a slender rod or wire of length L, cross-sectional area A, and density ρ, the body approximates a line segment and $\mathrm{d}m = \rho A\,\mathrm{d}L$, if ρ and A are constant over the length of the rod, the coordinates of the center of mass become the coordinates of the centroid C of the line segment, and (2.15) may be written

$$\overline{x} = \frac{\int x\,\mathrm{d}L}{L}\,, \quad \overline{y} = \frac{\int y\,\mathrm{d}L}{L}\,, \quad \overline{z} = \frac{\int z\,\mathrm{d}L}{L}\,. \tag{2.17}$$

Similarly, for area and volumes, the expressions of centroids are:

2. Areas

$$\overline{x} = \frac{\int x\,\mathrm{d}A}{A}\,, \quad \overline{y} = \frac{\int y\,\mathrm{d}A}{A}\,, \quad \overline{z} = \frac{\int z\,\mathrm{d}A}{A}\,. \tag{2.18}$$

3. Volumes

$$\overline{x} = \frac{\int x\,\mathrm{d}V}{V}\,, \quad \overline{y} = \frac{\int y\,\mathrm{d}V}{V}\,, \quad \overline{z} = \frac{\int z\,\mathrm{d}V}{V}\,, \tag{2.19}$$

where A and V represent the area and volume of the body, respectively. Some useful expressions of the centroid, area moment of inertia and mass moment of inertia for various geometric figures are listed in Tables 2.1, 2.2.

2.1.7 Friction

Consider a solid block resting on an unlubricated horizontal surface, with the application of a horizontal force $\boldsymbol{p}$ that continuously increases from zero to a value sufficient to move the block and give it an appreciable velocity (Fig. 2.14a). Note that the block does not move at first, which shows that a friction force $\boldsymbol{F}$ must have developed to balance $\boldsymbol{p}$. As the magnitude of $\boldsymbol{p}$ increases, the magnitude of $\boldsymbol{F}$ also increases until it reaches a maximum value $F_{\max} = F_{\mathrm{m}}$. If $\boldsymbol{p}$ is further increased, the block starts sliding and the magnitude of $\boldsymbol{F}$ drops from $\boldsymbol{F}_{\mathrm{m}}$ to a lower value $\boldsymbol{F}_{\mathrm{k}}$ (Fig. 2.14b). Experimental evidence shows that $\boldsymbol{F}_{\mathrm{m}}$ and $\boldsymbol{F}_{\mathrm{k}}$ are proportional to the normal component $\boldsymbol{N}$ of the reaction of

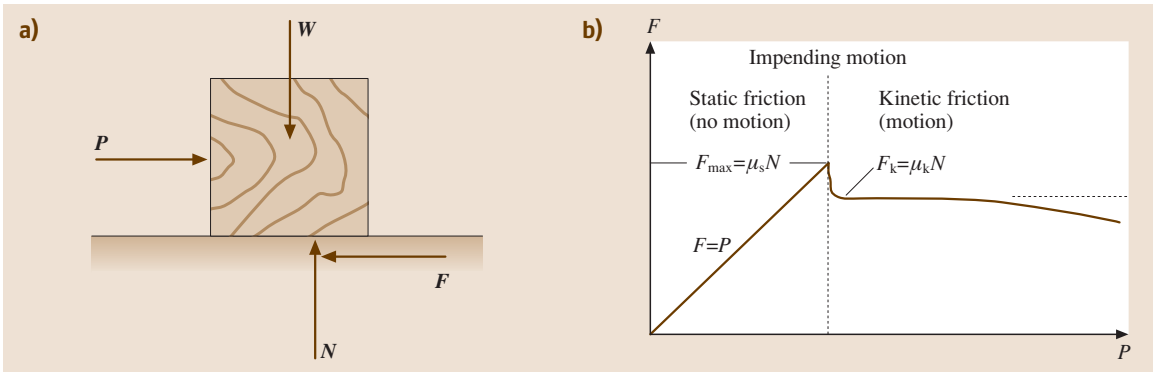


Fig. 2.14 (a) Dry friction, (b) relation of friction

Table 2.1 Properties of plane figures (after [2.2])

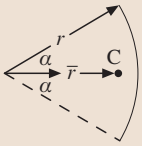
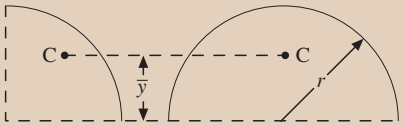
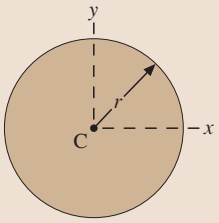
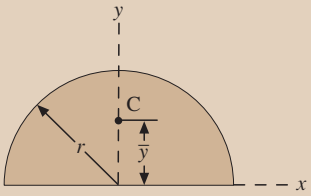
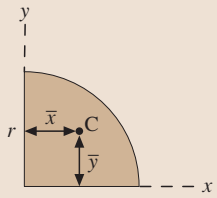
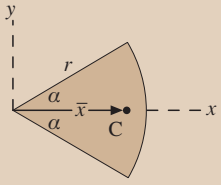
Figure	Centroid	Area moments of inertia
Arc segment 	$\overline{r} = \frac{r \sin\alpha}{\alpha}$	–
Quarter and semicircular arcs 	$\overline{y} = \frac{2r}{\pi}$	–
Circular area 	–	$I_x = I_y = \frac{\pi r^4}{4}$, $I_z = \frac{\pi r^4}{2}$
Semicircular area 	$\overline{y} = \frac{4r}{3\pi}$	$I_x = I_y = \frac{\pi r^4}{8}$, $\overline{I}_x = \left(\frac{\pi}{8} - \frac{8}{9\pi}\right) r^4$, $I_z = \frac{\pi r^4}{4}$
Quarter-circular area 	$\overline{x} = \overline{y} = \frac{4r}{3\pi}$	$I_x = I_y = \frac{\pi r^4}{16}$, $\overline{I}_x = \overline{I}_y = \left(\frac{\pi}{16} - \frac{4}{9\pi}\right) r^4$, $I_z = \frac{\pi r^4}{8}$
Area of circular sector 	$\overline{x} = \frac{2}{3} \frac{r \sin\alpha}{\alpha}$	$I_x = \frac{r^4}{4}\left(\alpha - \frac{1}{2}\sin 2\alpha\right)$, $I_y = \frac{r^4}{4}\left(\alpha + \frac{1}{2}\sin 2\alpha\right)$, $I_z = \frac{1}{2} r^4 \alpha$

Table 2.1 (cont.)

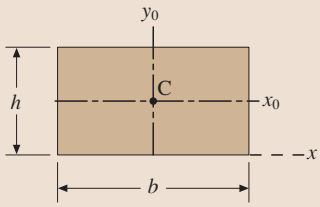
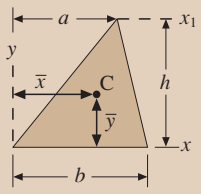
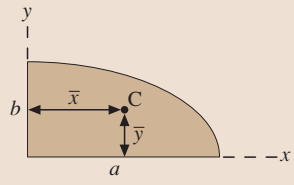
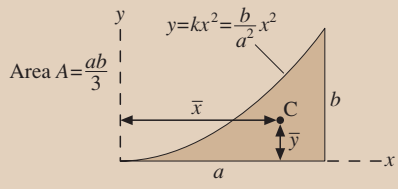
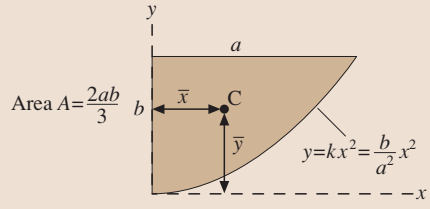
Figure	Centroid	Area moments of inertia
Rectangular area 	–	$I_x = \frac{bh^3}{3}$, $\overline{I}_x = \frac{bh^3}{12}$, $\overline{I}_z = \frac{bh}{12}(b^2+h^2)$
Triangular area 	$\overline{x} = \frac{a+b}{3}$, $\overline{y} = \frac{h}{3}$	$I_x = \frac{bh^3}{12}$, $\overline{I}_x = \frac{bh^3}{36}$, $I_{x_1} = \frac{bh^3}{4}$
Area of elliptical quadrant 	$\overline{x} = \frac{4a}{3\pi}$, $\overline{y} = \frac{4b}{3\pi}$	$I_x = \frac{\pi ab^3}{16}$, $\overline{I}_x = \left(\frac{\pi}{16} - \frac{4}{9\pi}\right)ab^3$, $I_y = \frac{\pi a^3 b}{16}$, $\overline{I}_y = \left(\frac{\pi}{16} - \frac{4}{9\pi}\right)a^3 b$, $I_z = \frac{\pi ab}{16}(a^2+b^2)$
Subparabolic area 	$\overline{x} = \frac{3a}{4}$, $\overline{y} = \frac{3b}{10}$	$I_x = \frac{ab^3}{21}$, $I_y = \frac{a^3 b}{5}$, $I_z = ab\left(\frac{a^2}{5} + \frac{b^2}{21}\right)$
Parabolic area 	$\overline{x} = \frac{3a}{8}$, $\overline{y} = \frac{3b}{5}$	$I_x = \frac{2ab^3}{7}$, $I_y = \frac{2a^3 b}{15}$, $I_z = 2ab\left(\frac{a^2}{15} + \frac{b^2}{7}\right)$

Table 2.2 Properties of homogeneous solids (m = mass of body shown)(after [2.2])

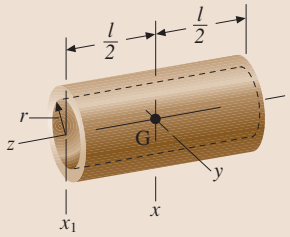
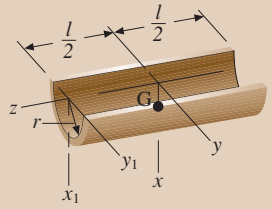
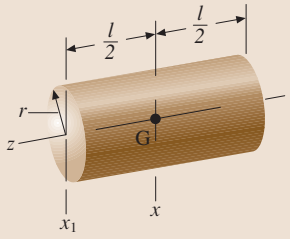
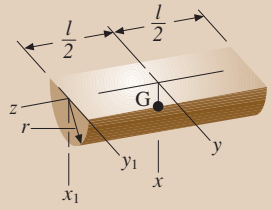
Body	Mass centre	Mass moment of inertia
Circular cylindric shell 	–	$I_{xx} = \frac{1}{2}mr^2 + \frac{1}{12}ml^2$, $I_{x_1x_1} = \frac{1}{2}mr^2 + \frac{1}{3}ml^2$, $I_{zz} = mr^2$
Half cylindric shell 	$\overline{x} = \frac{2r}{\pi}$	$I_{xx} = I_{yy} = \frac{1}{2}mr^2 + \frac{1}{12}ml^2$, $I_{x_1x_1} = I_{y_1y_1} = \frac{1}{2}mr^2 + \frac{1}{3}ml^2$, $I_{zz} = mr^2$, $\overline{I}_{zz} = \left(1 - \frac{4}{\pi^2}\right)mr^2$
Circular cylinder 	–	$I_{xx} = \frac{1}{4}mr^2 + \frac{1}{12}ml^2$, $I_{x_1x_1} = \frac{1}{4}mr^2 + \frac{1}{3}ml^2$, $I_{zz} = \frac{1}{2}mr^2$
Semicylinder 	$\overline{x} = \frac{4r}{3\pi}$	$I_{xx} = I_{yy} = \frac{1}{4}mr^2 + \frac{1}{12}ml^2$, $I_{x_1x_1} = I_{y_1y_1} = \frac{1}{4}mr^2 + \frac{1}{3}ml^2$ $I_{zz} = \frac{1}{2}mr^2$, $\overline{I}_{zz} = \left(\frac{1}{2} - \frac{16}{9\pi^2}\right)mr^2$

Table 2.2 (cont.)

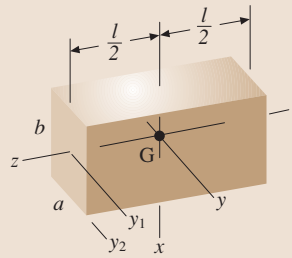
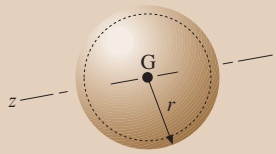
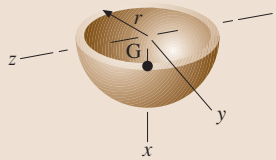
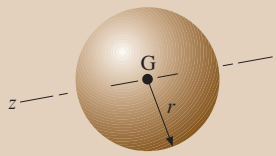
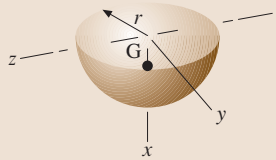
Body	Mass centre	Mass moment of inertia
Rectangular parallelpiped 	–	$I_{xx} = \frac{1}{12}m(a^2+l^2)$, $I_{yy} = \frac{1}{12}m(b^2+l^2)$, $I_{zz} = \frac{1}{12}m(a^2+b^2)$, $I_{y_1y_1} = \frac{1}{12}mb^2 + \frac{1}{3}ml^2$, $I_{y_2y_2} = \frac{1}{3}m(b^2+l^2)$
Spheric shell 	–	$I_{zz} = \frac{2}{3}mr^2$
Hemispherical shell 	$\overline{x} = \frac{r}{2}$	$I_{xx} = I_{yy} = I_{zz} = \frac{2}{3}mr^2$, $\overline{I}_{yy} = \overline{I}_{zz} = \frac{5}{12}mr^2$
Sphere 	–	$I_{zz} = \frac{2}{5}mr^2$
Hemisphere 	$\overline{x} = \frac{3r}{8}$	$I_{xx} = I_{yy} = I_{zz} = \frac{2}{5}mr^2$, $\overline{I}_{yy} = \overline{I}_{zz} = \frac{83}{320}mr^2$

Table 2.2 (cont.)

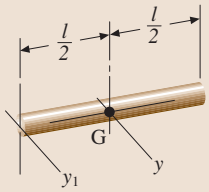
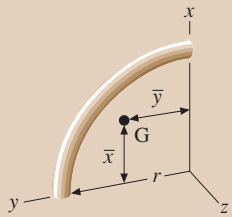
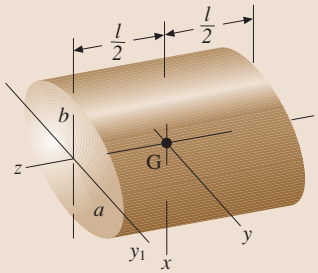
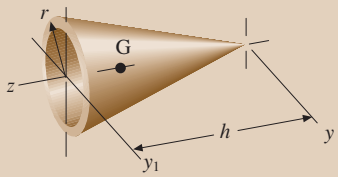
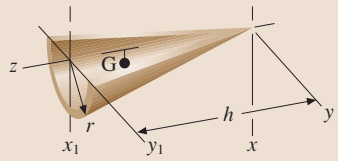
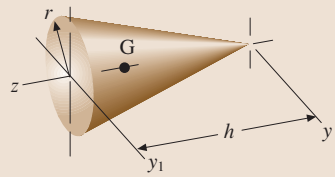
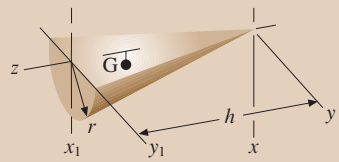
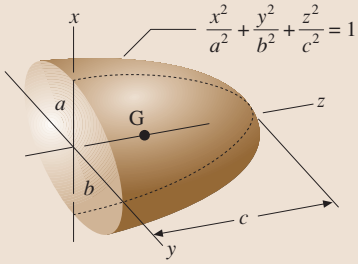
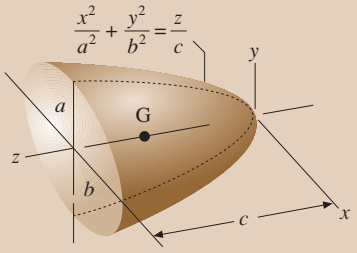
Body	Mass centre	Mass moment of inertia
Uniform slender rod 	–	$I_{yy} = \frac{1}{12}ml^2\,, \quad I_{y_1y_1} = \frac{1}{3}ml^2$
Quarter circular rod 	$\overline{x} = \overline{y} = \frac{2r}{\pi}$	$I_{xx} = I_{yy} = \frac{1}{2}mr^2\,, \quad I_{zz} = mr^2$
Elliptical cylinder 	–	$I_{xx} = \frac{1}{4}ma^2 + \frac{1}{12}ml^2\,, \quad I_{yy} = \frac{1}{4}mb^2 + \frac{1}{12}ml^2$ $I_{zz} = \frac{1}{4}m(a^2+b^2)\,, \quad I_{y_1y_1} = \frac{1}{4}mb^2 + \frac{1}{3}ml^2$
Conical shell 	$\overline{z} = \frac{2h}{3}$	$I_{yy} = \frac{1}{4}mr^2 + \frac{1}{2}mh^2\,, \quad I_{y_1y_1} = \frac{1}{4}mr^2 + \frac{1}{6}mh^2$ $I_{zz} = \frac{1}{2}mr^2\,, \quad \overline{I}_{zz} = \frac{1}{4}mr^2 + \frac{1}{18}mh^2$
Half conical shell 	$\overline{x} = \frac{4r}{3\pi}\,, \quad \overline{z} = \frac{2h}{3}$	$I_{xx} = I_{yy} = \frac{1}{4}mr^2 + \frac{1}{2}mh^2\,, \quad I_{x_1x_1} = I_{y_1y_1} = \frac{1}{4}mr^2 + \frac{1}{6}mh^2$ $I_{zz} = \frac{1}{2}mr^2\,, \quad \overline{I}_{zz} = \left(\frac{1}{2} - \frac{16}{9\pi^2}\right)mr^2$

Table 2.2 (cont.)

Body	Mass centre	Mass moment of inertia
Right circular cone 	$\overline{z} = \frac{3h}{4}$	$I_{yy} = \frac{3}{20}mr^2 + \frac{3}{5}mh^2\,, \quad I_{y_1y_1} = \frac{3}{20}mr^2 + \frac{1}{10}mh^2$ $I_{zz} = \frac{3}{10}mr^2\,, \quad \overline{I}_{yy} = \frac{3}{20}mr^2 + \frac{3}{80}mh^2$
Half cone 	$\overline{x} = \frac{r}{\pi}\,, \quad \overline{z} = \frac{3h}{4}$	$I_{xx} = I_{yy} = \frac{3}{20}mr^2 + \frac{3}{5}mh^2$ $I_{x_1x_1} = I_{y_1y_1} = \frac{3}{20}mr^2 + \frac{1}{10}mh^2$ $I_{zz} = \frac{3}{10}mr^2\,, \quad \overline{I}_{zz} = (\frac{3}{10}mr^2 - \frac{1}{\pi^2})mr^2$
Semiellipsoid 	$\overline{z} = \frac{3c}{8}$	$I_{xx} = \frac{1}{5}m(b^2 + c^2)\,, \quad I_{yy} = \frac{1}{5}m(a^2 + c^2)$ $I_{zz} = \frac{1}{5}m(a^2 + b^2)\,, \quad \overline{I}_{xx} = \frac{1}{5}m(b^2 + \frac{19}{64}c^2)$ $\overline{I}_{yy} = \frac{1}{5}m(a^2 + \frac{19}{64}c^2)$
Eliptic paraboloid 	$\overline{z} = \frac{2c}{3}$	$I_{xx} = \frac{1}{6}mb^2 + \frac{1}{2}mc^2\,, \quad I_{yy} = \frac{1}{6}ma^2 + \frac{1}{2}mc^2$ $I_{zz} = \frac{1}{6}m(a^2 + b^2)\,, \quad \overline{I}_{xx} = \frac{1}{6}m(b^2 + \frac{1}{3}c^2)$ $\overline{I}_{yy} = \frac{1}{6}m(a^2 + \frac{1}{3}c^2)$

the surface. Thus

$$F_m = \mu_s N\,, \quad F_k = \mu_k N\,, \tag{2.20}$$

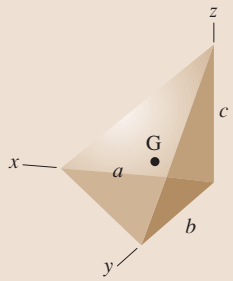
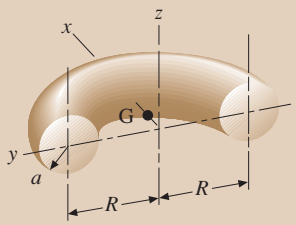
where μ_s and μ_k are called, respectively, the coefficients of static and kinetic friction.

It is convenient to replace the normal force $\boldsymbol{N}$ and the friction force $\boldsymbol{F}$ by their resultant $\boldsymbol{R}$, as shown in Fig. 2.15. As the friction force increases and reaches its maximum value $F_m = \mu_s N$, the angle ϕ that $\boldsymbol{R}$ forms with the normal to the surface increases and reaches a maximum value ϕ_s, called the angle of static friction. If motion actually takes place, the magnitude of $\boldsymbol{F}$ drops to $\boldsymbol{F}_k$, similarly the angle ϕ drops to a lower value ϕ_k, called the angle of kinetic friction. Thus

$$\tan\phi_s = \mu_s\,, \quad \tan\phi_k = \mu_k\,. \tag{2.21}$$

Dry friction plays an important role in a number of engineering applications such as wedges, square-threaded screws, journal bearings, thrust bearings, and disk friction. In solving a problem involving a flat belt passing over a fixed cylinder, it is important first to determine the direction in which the belt slips or is about

Table 2.2 (cont.)

Body	Mass centre	Mass moment of inertia
Rectangular thetrahedron 	$\overline{x} = \frac{a}{4}$, $\overline{y} = \frac{b}{4}$ $\overline{z} = \frac{c}{4}$	$I_{xx} = \frac{1}{10}m(b^2+c^2)$, $I_{yy} = \frac{1}{10}m(a^2+c^2)$ $I_{zz} = \frac{1}{10}m(a^2+b^2)$, $\overline{I}_{xx} = \frac{3}{80}m(b^2+c^2)$ $\overline{I}_{yy} = \frac{3}{80}m(a^2+c^2)$, $\overline{I}_{zz} = \frac{3}{80}m(a^2+b^2)$
Half torus 	$\overline{x} = \frac{a^2+4R^2}{2\pi R}$	$I_{xx} = I_{yy} = \frac{1}{2}mR^2 + \frac{5}{8}ma^2$, $I_{zz} = mR^2 + \frac{3}{4}ma^2$

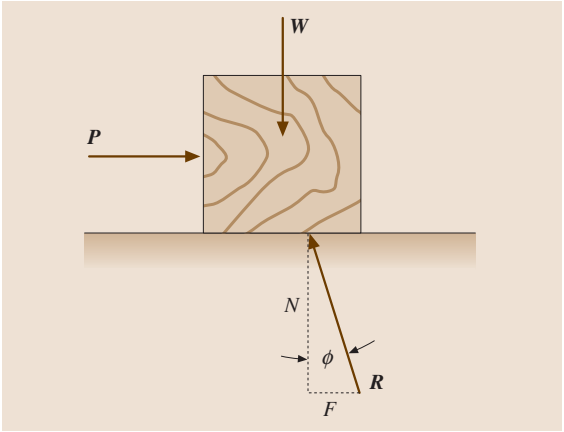


Fig. 2.15 Free-body diagram of a solid block

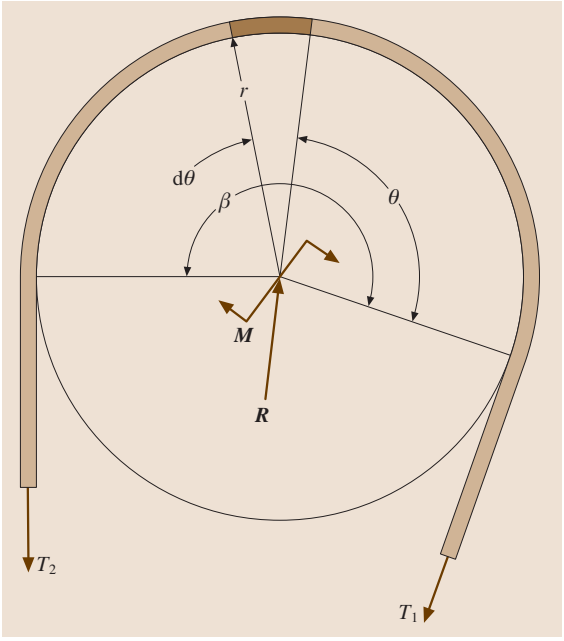


Fig. 2.16 A drum subjected to the two belt tensions T_1 and T_2 and the torque $\boldsymbol{M}$

to slip. If the drum is rotating, the motion or impending motion of the belt should be determined relative to the rotating drum. For instance, if the belt shown in Fig. 2.16 with M in the direction shown, T_2 is greater than T_1. Denoting the larger tension by T_2, the smaller tension by T_1, the coefficient of static friction by μ_s, and the angle (in radius) subtended by the belt by β, the

following formulas can be derived

$$\ln \frac{T_2}{T_1} = \mu_s \beta \,, \tag{2.22}$$

$$\frac{T_2}{T_1} = e^{\mu_s \beta} \,. \tag{2.23}$$

If the belt actually slips on the drum, the coefficient of static friction μ_s should be replaced by the coefficient of kinetic friction μ_k in both of these formulas.

It is important to note that a coefficient of friction applies to a given pair of mating surfaces. It is meaningless to speak of a coefficient of friction for a single surface. Also friction coefficients vary considerably, depending on the exact condition of the mating surfaces.

2.1.8 Principle of Virtual Work

The principle of virtual work for a particle states that *if a particle is in equilibrium, the total virtual work done by the n applied forces during any arbitrary virtual displacement of the particle is zero*. This can easily be verified as follows.

Let the virtual displacement be $\delta \boldsymbol{r}$, then the virtual work done by any force $\boldsymbol{F}_i$ $(i = 1, 2, \ldots, n)$ is the product of the virtual displacement and the component of the force in the direction of the virtual displacement, thus

$$\begin{aligned} \delta U &= \boldsymbol{F}_1 \cdot \delta \boldsymbol{r} + \boldsymbol{F}_2 \cdot \delta \boldsymbol{r} + \cdots + \boldsymbol{F}_n \cdot \delta \boldsymbol{r} \\ &= \left(\sum \boldsymbol{F} \right) \cdot \delta \boldsymbol{r} \,. \end{aligned}$$

However, the equilibrium of these static forces requires that the sum of these forces in any direction be zero, hence

$$\delta U = 0 \,.$$

That is, the total virtual work done during any virtual displacement is zero.

In the case of a rigid body, the principle of virtual work states that: *if a rigid body is in equilibrium, the total virtual work of the external forces acting on the rigid body is zero for any virtual displacement of the body*.

The principle of virtual work can be extended to the case of a system of connected rigid bodies. If the system remains connected during a virtual displacement, only the work of the forces external to the system need be considered, since the total work of the internal forces at the various connections is zero.

2.2 Dynamics

Dynamic is one of the oldest branches of physics, with its development as a science beginning with Galileo about four centuries ago. His experiments on uniformly accelerated bodies led Newton to formulate his fundamental laws of motion. The study of dynamics of particles and rigid bodies as an engineering subject is not so old, perhaps going back to after World War II as a standard course in engineering curricula. An even later addition to engineering curricula is the study of vibrations, which can be regarded as the part of dynamics concerned with the motion of elastic systems.

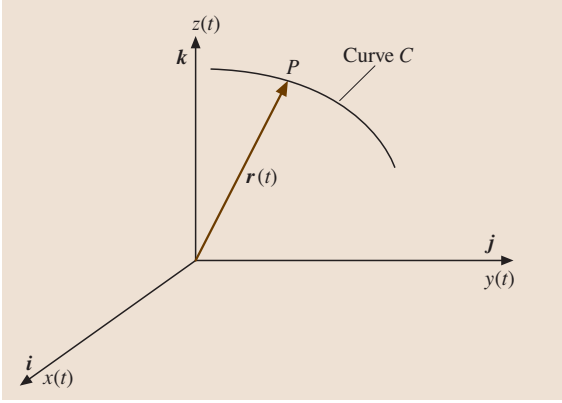


Fig. 2.17 The motion of a particle

The study of the motion of a body without regard to the forces and moments causing the motion is known as kinematics. One may think of kinematics as the geometry of motion. The material presented here is fundamental to the dynamics of systems of particles and rigid bodies [2.5–8].

2.2.1 Motion of a Particle

Motion Relative to a Fixed Frame

The position of a particle P in space is defined at any time t by the three Cartesian coordinates $x(t)$, $y(t)$, and $z(t)$. To describe the motion of a particle P along curve C in a three-dimensional space, as depicted in Fig. 2.17,

the position vector is expressed as

$$\boldsymbol{r}(t) = x(t)\boldsymbol{i} + y(t)\boldsymbol{j} + z(t)\boldsymbol{k} \,. \tag{2.24}$$

Velocity

The velocity of the particle P in space is defined as the time rate of change of the position. The velocity vector of P is written as

$$\begin{aligned}\boldsymbol{v}(t) &= \frac{\mathrm{d}\boldsymbol{r}(t)}{\mathrm{d}t} = \dot{x}(t)\boldsymbol{i} + \dot{y}(t)\boldsymbol{j} + \dot{z}(t)\boldsymbol{k} \\ &= v_x(t)\boldsymbol{i} + v_y(t)\boldsymbol{j} + v_z(t)\boldsymbol{k} \,,\end{aligned} \tag{2.25}$$

where

$$v_x(t) = \dot{x}(t) \,, \quad v_y(t) = \dot{y}(t) \,, \quad v_z(t) = \dot{z}(t) \tag{2.26}$$

are the Cartesian components of the velocity vector. From Fig. 2.17, as $\Delta t \to 0$, the increment $\Delta\boldsymbol{r}(t)$ in the position vector corresponding to the time increment aligns itself with the curve C and becomes the differential $\mathrm{d}\boldsymbol{r}(t)$. Hence, the velocity vector is tangent to the curve trajectory C at all time.

Acceleration

The acceleration of the particle P in space is defined as the time rate of change of the velocity. The acceleration vector of P is written as

$$\begin{aligned}\boldsymbol{a}(t) &= \frac{\mathrm{d}\boldsymbol{v}(t)}{\mathrm{d}t} = \dot{v}_x(t)\boldsymbol{i} + \dot{v}_y(t)\boldsymbol{j} + \dot{v}_z(t)\boldsymbol{k} \\ &= a_x(t)\boldsymbol{i} + a_y(t)\boldsymbol{j} + a_z(t)\boldsymbol{k} \,,\end{aligned} \tag{2.27}$$

where

$$\begin{aligned}&a_x(t) = \dot{v}_x(t) = \ddot{x}(t) \,, \quad a_y(t) = \dot{v}_y(t) = \ddot{y}(t) \,, \\ &a_z(t) = \dot{v}_z(t) = \ddot{z}(t)\end{aligned} \tag{2.28}$$

are the Cartesian components of the acceleration vector.

Rectilinear Motion

Rectilinear motion implies motion along a straight line. Since there is only one component of motion, we may dispense with the vector notation and describe the motion in terms of scalar quantities. Denoting the line along which the motion takes place by s and the distance of the particle P from the fixed origin 0 by $s(t)$ as depicted in Fig. 2.18, the velocity of P is written as

$$v(t) = \frac{\mathrm{d}s(t)}{\mathrm{d}t} = \dot{s}(t) \,. \tag{2.29}$$

The acceleration of P is written as

$$a(t) = \frac{\mathrm{d}v(t)}{\mathrm{d}t} = \frac{\mathrm{d}^2 s(t)}{\mathrm{d}t^2} = \ddot{s}(t) \,. \tag{2.30}$$

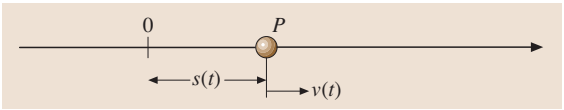


Fig. 2.18 Rectilinear motion

The distance $s(t)$, the velocity $v(t)$, the acceleration $a(t)$ are explicit function of t. However, one can derive a relation among s, v, and a in which the time t is only implicit.

Let us use the chain rule for differentiation and write

$$a = \frac{\mathrm{d}v}{\mathrm{d}t} = \frac{\mathrm{d}v}{\mathrm{d}s}\frac{\mathrm{d}s}{\mathrm{d}t} = \frac{\mathrm{d}v}{\mathrm{d}s}v \tag{2.31}$$

or

$$a\,\mathrm{d}s = v\,\mathrm{d}v \,. \tag{2.32}$$

Integrating (2.32) between the points $s = s_1$, $v = v_1$ and $s = s_2$, $v = v_2$, we obtain

$$\int_{s_1}^{s_2} a\,\mathrm{d}s = \int_{v_1}^{v_2} v\,\mathrm{d}v = \frac{1}{2}\left(v_2^2 - v_1^2\right) . \tag{2.33}$$

Uniform Motion

In this case, $v(t) = v_0$= constant. By integration, it follows that

$$s(t) = \int v\,\mathrm{d}t = v_0 t + C_1 \,. \tag{2.34}$$

With the initial condition, $s(t = t_1) = s_1$, from this $C_1 = s_1 - v_0 t_1$, therefore,

$$s(t) = v_0(t - t_1) + s_1 \,. \tag{2.35}$$

Uniform Accelerated Motion

In this case, $a(t) = a_0 =$ constant. By integration, it follows that

$$v(t) = \int a\,\mathrm{d}t = a_0 t + C_1 \tag{2.36}$$

or

$$s(t) = \int v\,\mathrm{d}t = a_0\frac{t^2}{2} + C_1 t + C_2 \,. \tag{2.37}$$

With the initial condition, $v(t = t_1) = v_1$, and $s(t = t_1) = s_1$, the constants follow: $C_1 = v_1 - a_0 t_1$ and $C_2 = s_1 - v_1 t_1 + a_0 t_1^2/2$ and therefore

$$s(t) = a_0\frac{(t - t_1)^2}{2} + v_1(t - t_1) + s_1 \,. \tag{2.38}$$

Example 2.1: A car starting from rest travels with constant acceleration a_0 for 10 s. Determine the value a_0 given that the car has reached a velocity of 108 km/h at the end of the 5 s. What is the distance traveled by the car?

$$v(t) = a_0 t$$

$$a_0 = \left.\frac{v(t)}{5}\right|_{t=5} = \frac{108(1000)/3600}{5} = 6\,\mathrm{m/s^2}$$

$$s = \frac{1}{2}a_0 t^2 = \frac{1}{2}6(5)^2 = 75\,\mathrm{m}$$

Nonuniform Accelerated Motion

In this case, $a(t) = f_0(t)$. By integration, it follows that

$$v(t) = \int a(t)\,\mathrm{d}t = \int f_0(t)\,\mathrm{d}t = f_1(t) + C_1 \qquad (2.39)$$

or

$$s(t) = \int v(t)\,\mathrm{d}t = \int [f_1(t) + C_1]\,\mathrm{d}t$$

$$= f_2(t) + C_1 t + C_2\,. \qquad (2.40)$$

The constants are determined from the initial conditions or equivalent conditions.

2.2.2 Planar Motion, Trajectories

Consider a particle traveling in the xz plane with constant acceleration (gravity) $\ddot{z} = -g$ after the initial velocity v_0 as shown in Fig. 2.19. Let v_0 be the magnitude of the initial velocity and α the angle between v_0 and the x-axis; the velocity of the particle at time t in Cartesian components is

$$\boldsymbol{v} = v_0 \cos\alpha\,\boldsymbol{i} + (v_0 \sin\alpha - gt)\boldsymbol{k}\,. \qquad (2.41)$$

The trajectory of the particle at time t in Cartesian components is

$$x = v_0 t \cos\alpha\,, \quad z = v_0 t \sin\alpha - \frac{1}{2}gt^2\,, \qquad (2.42)$$

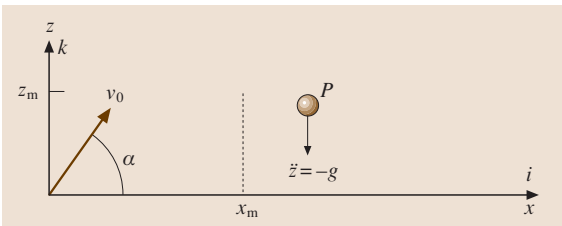


Fig. 2.19 Planar motion

from which the travel time and trajectory can then computed as

$$t = \frac{x}{v_0 \cos\alpha}\,, \quad z = x\tan\alpha - \frac{1}{2}\frac{gx^2}{v_0^2\cos^2\alpha}\,. \qquad (2.43)$$

This trajectory represents a parabola. To determine its shape, its maximum altitude is calculated by locating the point with zero slope

$$\frac{\mathrm{d}z}{\mathrm{d}x} = \tan\alpha - \frac{gx}{v_0^2\cos^2\alpha} = 0\,. \qquad (2.44)$$

Let x_m be the distance along the x-axis corresponding to the maximum altitude z_m; one obtains

$$x_m = \frac{v_0^2 \sin\alpha\cos\alpha}{g}\,. \qquad (2.45)$$

The maximum altitude z_m is

$$z_m = \frac{v_0^2}{2g}\sin^2\alpha\,. \qquad (2.46)$$

The trajectory is symmetrical with respect to the vertical through x_m, as shown in Fig. 2.19. One concludes that the particle hits the ground at $x_f = 2x_m$. The final velocity $\boldsymbol{v}_f$ in Cartesian component is

$$v_x = v_0\cos\alpha\,, \quad v_z = -v_0\sin\alpha\,. \qquad (2.47)$$

2.2.3 Polar Coordinates

Consider a particle traveling along curve C as shown in Fig. 2.20. In polar coordinates, one defines the radial axis r as the axes coinciding at all times with the direction of the radius vector $\boldsymbol{r}(t)$ from the origin O to the point P. The transverse axis θ is normal to the radial axis as shown in Fig. 2.20. The unit vectors $\boldsymbol{u}_r(t)$

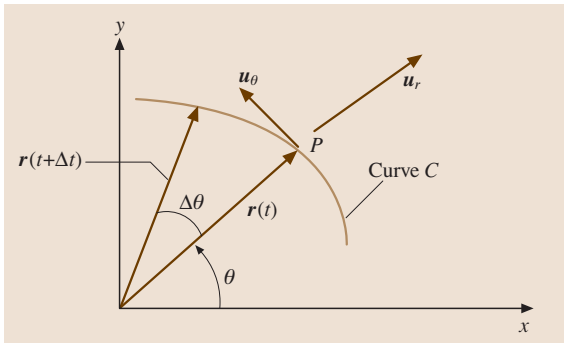


Fig. 2.20 Polar coordinates

and $\boldsymbol{u}_\theta(t)$, representing the radial and transverse directions, are functions of time. This can be explained by observing that the radius vector $\boldsymbol{r}(t)$ changes directions continuously as the particle moves along the curve C. Because the unit vector $\boldsymbol{u}_r(t)$ is aligned with $\boldsymbol{r}(t)$, $\boldsymbol{u}_r(t)$ also changes direction continuously. Because $\boldsymbol{u}_\theta(t)$ is normal to $\boldsymbol{u}_r(t)$, $\boldsymbol{u}_\theta(t)$ is also a time-dependent unit vector.

The location of the particle P is expressed as

$$\boldsymbol{r}(t) = r(t)\boldsymbol{u}_r(t) \tag{2.48}$$

and the velocity of P is

$$\boldsymbol{v}(t) = \dot{\boldsymbol{r}}(t) = \dot{r}(t)\boldsymbol{u}_r(t) + r(t)\dot{\boldsymbol{u}}_r(t)\,. \tag{2.49}$$

From Fig. 2.21, it follows that

$$\begin{aligned}\dot{\boldsymbol{u}}_r(t) &= \lim_{\Delta t\to 0}\frac{\Delta \boldsymbol{u}_r(t)}{\Delta t} = \lim_{\Delta t\to 0}\frac{\Delta\theta(t)}{\Delta t}\boldsymbol{u}_\theta(t)\\ &= \dot{\theta}(t)\boldsymbol{u}_\theta(t)\end{aligned} \tag{2.50}$$

$$\begin{aligned}\dot{\boldsymbol{u}}_\theta(t) &= \lim_{\Delta t\to 0}\frac{\Delta \boldsymbol{u}_\theta(t)}{\Delta t} = \lim_{\Delta t\to 0}\frac{\Delta\theta(t)}{\Delta t}[-\boldsymbol{u}_r(t)]\\ &= -\dot{\theta}(t)\boldsymbol{u}_r(t)\,.\end{aligned} \tag{2.51}$$

The velocity of P is rewritten as

$$\begin{aligned}\boldsymbol{v}(t) &= \dot{\boldsymbol{r}}(t) = \dot{r}(t)\boldsymbol{u}_r(t) + r(t)\dot{\boldsymbol{u}}_r(t)\\ &= v_r\boldsymbol{u}_r(t) + v_\theta\boldsymbol{u}_\theta(t)\,,\end{aligned} \tag{2.52}$$

where $v_r = \dot{r}$ and $v_\theta = r\dot{\theta}$ are the radial and transverse components of the velocity vector, respectively.

Similarly, the acceleration of point P is

$$\begin{aligned}\boldsymbol{a}(t) &= \dot{\boldsymbol{v}}(t) = \ddot{\boldsymbol{r}}(t)\\ &= \ddot{r}\boldsymbol{u}_r(t) + \dot{r}(t)\dot{\boldsymbol{u}}_r(t) + \dot{r}\dot{\theta}\boldsymbol{u}_\theta + r\ddot{\theta}\boldsymbol{u}_\theta + r\dot{\theta}\dot{\boldsymbol{u}}_\theta\\ &= (\ddot{r} - r\,\dot{\theta}^2)\boldsymbol{u}_r + (r\ddot{\theta} + 2\dot{r}\dot{\theta})\boldsymbol{u}_\theta = a_r\boldsymbol{u}_r + a_\theta\boldsymbol{u}_\theta\,,\end{aligned} \tag{2.53a}$$

where

$$a_r = \ddot{r} - r\dot{\theta}^2 \quad \text{and} \quad a_\theta = r\ddot{\theta} + 2\dot{r}\dot{\theta} \tag{2.53b}$$

are the radial and transverse components of the acceleration vector, respectively.

It should be noted that, by adding the coordinate z to the polar coordinates r and θ, one obtains the cylindrical coordinates, r, θ, z. The velocity and acceleration vectors can be expressed in terms of cylindrical coordinates as

$$\boldsymbol{v}(t) = v_r\boldsymbol{u}_r + v_\theta\boldsymbol{u}_\theta + v_z\boldsymbol{k}\,, \tag{2.54}$$

where $v_z = \dot{z}$,

$$\boldsymbol{a}(t) = a_r\boldsymbol{u}_r + a_\theta\boldsymbol{u}_\theta + a_z\boldsymbol{k}\,, \tag{2.55}$$

and $a_z = \ddot{z}$.

Example 2.2: A bicycler enters a semicircular track of radius $r = 60\,\mathrm{m}$, as shown in Fig. 2.22, with velocity $v_\mathrm{A} = 18\,\mathrm{m/s}$, decelerates at a uniform rate, and exits with velocity $v_\mathrm{C} = 12\,\mathrm{m/s}$. Find the circumferential deceleration, the time it takes to complete the semicircle, and the velocity at point B.

Denote the magnitude of the circumferential deceleration as follows

$$a_\theta = r\ddot{\theta}\,. \tag{2.56a}$$

Integrating (2.56a) with respect to time yields

$$v_\theta = r\dot{\theta} = v_\mathrm{A} + a_\theta t\,. \tag{2.56b}$$

Letting $t = t_\mathrm{f}$ and $v_\theta = v_\mathrm{C}$ in (2.56b), one can write

$$a_\theta t_\mathrm{f} = v_\mathrm{C} - v_\mathrm{A} = 12 - 18 = -6\,\mathrm{m/s}\,. \tag{2.56c}$$

Integrating (2.56b) with respect to time, one obtains

$$r\theta = v_\mathrm{A}t + \frac{1}{2}a_\theta t^2\,. \tag{2.56d}$$

Inserting $t = t_\mathrm{f}$, $a_\theta t_\mathrm{f} = -6$, and $\theta_\mathrm{f} = \pi$ in (2.56d) yields

$$t_\mathrm{f} = \frac{r\theta_\mathrm{f}}{v_\mathrm{A} + \frac{1}{2}a_\theta t_\mathrm{f}} = \frac{60(\pi)}{18 + \frac{1}{2}(-6)} = 4\pi\,\mathrm{s} \tag{2.56e}$$

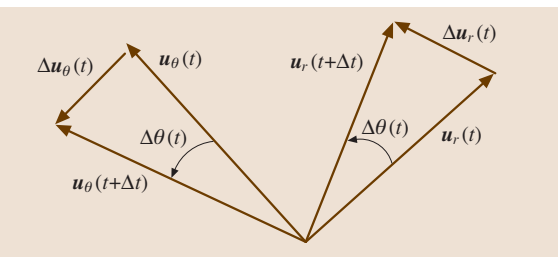


Fig. 2.21 Motion in polar coordinates

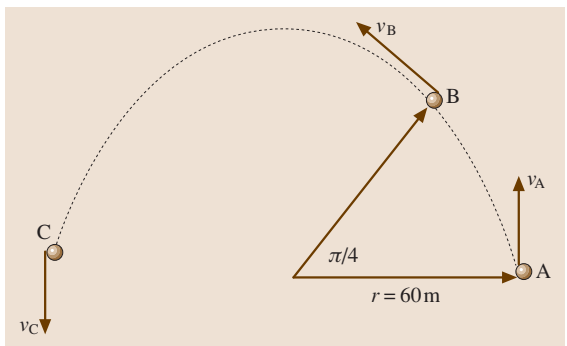


Fig. 2.22 A bicycler in the semicircular track

and

$$a_\theta = \frac{-6}{4\pi}\,\mathrm{m/s^2}\,. \tag{2.56f}$$

To obtain v_B, we use

$$a_\theta s_{AB} = \frac{1}{2}\left(v_B^2 - v_A^2\right) \tag{2.56g}$$

so that

$$v_B = \sqrt{v_A^2 + 2a_\theta s_{AB}} = \sqrt{18^2 + 2\left(\frac{-6}{4\pi}\right)\left(\frac{60\pi}{4}\right)}$$
$$= 16.7\,\mathrm{m/s}\,. \tag{2.56h}$$

2.2.4 Motion of Rigid Bodies (Moving Reference Frames)

Consider a reference frame xyz moving relative to the fixed reference frame XYZ as shown in Fig. 2.23. A point P relative to the system XYZ is expressed as

$$\boldsymbol{R} = \boldsymbol{r}_A + \boldsymbol{r}_{AP}\,. \tag{2.57}$$

The velocity of P relative to an inertial space is

$$\boldsymbol{v} = \dot{\boldsymbol{R}} = \boldsymbol{v}_A + \boldsymbol{v}_{AP}\,, \tag{2.58}$$

where

$$\boldsymbol{v}_A = \dot{\boldsymbol{r}}_A \quad \text{and} \quad \boldsymbol{v}_{AP} = \dot{\boldsymbol{r}}_{AP}\,. \tag{2.59}$$

Similarly, the acceleration of P relative to an inertial space is

$$\boldsymbol{a} = \dot{\boldsymbol{v}} = \ddot{\boldsymbol{R}} = \boldsymbol{a}_A + \boldsymbol{a}_{AP}\,, \tag{2.60}$$

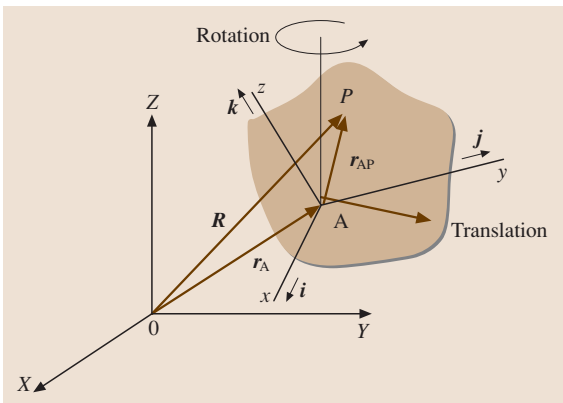


Fig. 2.23 Moving reference frame

where

$$\boldsymbol{a}_A = \dot{\boldsymbol{v}}_A = \ddot{\boldsymbol{r}}_A \quad \text{and} \quad \boldsymbol{a}_{AP} = \dot{\boldsymbol{v}}_{AP} = \ddot{\boldsymbol{r}}_{AP}\,. \tag{2.61}$$

Rotating Reference Frames

Assume that the rigid body is rotating about the axis AB, and consider a point P at a distance $d = |\boldsymbol{d}|$ from point C on axis AB, where C is the intersection between the axis AB and a plane normal to AB that contains the point P, as shown in Fig. 2.24. In the time increment Δt the vector $\boldsymbol{d}$ from C to P sweeps an angle $\Delta\theta$ in a plane normal to AB. The vector $\boldsymbol{d}$ rotates in a plane normal to the axis AB with the angular rate $\dot{\theta}$. Because the vector $\boldsymbol{d}$ is embedded in the rigid body, we say that the rigid body and the triad xyz rotate about the axis AB at the same rate $\dot{\theta}$. This angular rate is represented as a vector, ω, and is directed along the axis AB. Note that ω can be used to represent the angular velocity of the rigid body, or of the frame xyz, with units of radians per second (rad/s). We write

$$\lim_{\Delta t \to 0} \frac{\Delta\theta}{\Delta t} = \dot{\theta} = |\boldsymbol{\omega}|\,. \tag{2.62}$$

Consider the rate of change $\dot{\boldsymbol{d}}$ of the vector $\boldsymbol{d}$ due to the rotation of the body. From Fig. 2.24, we observe that the tip P of the vector $\boldsymbol{d}$ describes a circle of radius d, so that $\dot{\boldsymbol{d}}$ is tangent to the circle at P and is normal to the plane defined by the vectors $\boldsymbol{\omega}$ and $\boldsymbol{d}$. In the time increment Δt, the vector $\boldsymbol{d}$ makes the change in magnitude

$$\Delta d = d\Delta\theta \tag{2.63}$$

and

$$\dot{d} = \lim_{\Delta t \to 0} \frac{\Delta d}{\Delta t} = d\dot{\theta}\,. \tag{2.64}$$

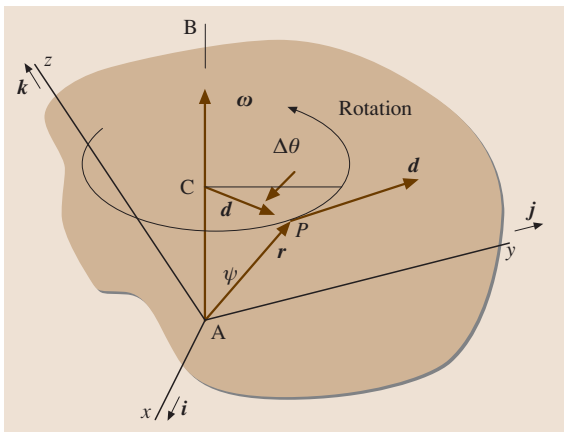


Fig. 2.24 Rotating rigid body

In vector, one has

$$\dot{\boldsymbol{d}} = \boldsymbol{\omega} \times \boldsymbol{d} \,. \tag{2.65}$$

One can generalize the above equation by observing that

$$d = r \, \sin\psi \,, \tag{2.66}$$

where r is the magnitude of the vector $\boldsymbol{r}$ and ψ is the angle between AB and $\boldsymbol{r}$. Hence the rate of change of the vector $\boldsymbol{r}$ is

$$\dot{\boldsymbol{r}} = \boldsymbol{\omega} \times \boldsymbol{r} \,. \tag{2.67}$$

The velocity of point P relative to point A is due entirely to the rotation of the frame xyz. Replacing the vector $\boldsymbol{r}$ by the radius vector $\boldsymbol{r}_{\mathrm{AP}}$, yields

$$\boldsymbol{v}_{\mathrm{AP}} = \dot{\boldsymbol{r}}_{\mathrm{AP}} = \boldsymbol{\omega} \times \boldsymbol{r}_{\mathrm{AP}} \,. \tag{2.68}$$

Because point A is at rest, $\boldsymbol{v}_{\mathrm{AP}}$ is also the absolute velocity of P, $\boldsymbol{v} = \boldsymbol{v}_{\mathrm{AP}}$. The time rate of change of $\boldsymbol{\omega}$ is referred as the angular acceleration with units of rad/s^2

$$\boldsymbol{\alpha} = \dot{\boldsymbol{\omega}} \,. \tag{2.69}$$

The acceleration of point P relative to point A is

$$\begin{aligned} \boldsymbol{a}_{\mathrm{AP}} &= \dot{\boldsymbol{v}}_{\mathrm{AP}} = \dot{\boldsymbol{\omega}} \times \boldsymbol{r}_{\mathrm{AP}} + \boldsymbol{\omega} \times \dot{\boldsymbol{r}}_{\mathrm{AP}} \\ &= \boldsymbol{\alpha} \times \boldsymbol{r}_{\mathrm{AP}} + \boldsymbol{\omega} \times (\boldsymbol{\omega} \times \boldsymbol{r}_{\mathrm{AP}}) \,. \end{aligned} \tag{2.70}$$

Because point A is at rest, $\boldsymbol{a}_{\mathrm{AP}}$ is also the absolute acceleration of P, $\boldsymbol{a} = \boldsymbol{a}_{\mathrm{AP}}$.

When the origin A of the rotating frame xyz is not fixed, but moves relative to the inertial frame XYZ with the velocity $\boldsymbol{v}_{\mathrm{A}}$ and acceleration $\boldsymbol{a}_{\mathrm{A}}$, then the absolute velocity and acceleration of P are

$$\boldsymbol{v} = \boldsymbol{v}_{\mathrm{A}} + \boldsymbol{v}_{\mathrm{AP}} = \boldsymbol{v}_{\mathrm{A}} + \boldsymbol{\omega} \times \boldsymbol{r}_{\mathrm{AP}} \,, \tag{2.71}$$

$$\boldsymbol{a} = \boldsymbol{a}_{\mathrm{A}} + \boldsymbol{a}_{\mathrm{AP}} = \boldsymbol{a}_{\mathrm{A}} + \boldsymbol{\alpha} \times \boldsymbol{r}_{\mathrm{AP}} + \boldsymbol{\omega} \times (\boldsymbol{\omega} \times \boldsymbol{r}_{\mathrm{AP}}) \,. \tag{2.72}$$

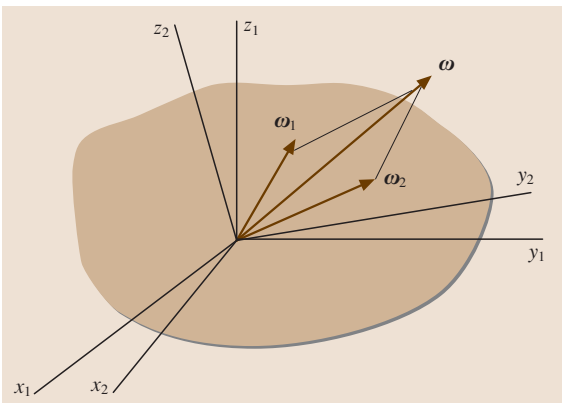


Fig. 2.25 Rotating frame

Consider the case in which the reference frame $x_1y_1z_1$ rotates with angular velocity $\boldsymbol{\omega}_1$ and the reference frame $x_2y_2z_2$ rotates relative to the frame $x_1y_1z_1$ with angular velocity $\boldsymbol{\omega}_2$, as shown in Fig. 2.25. The angular velocity of the frame $x_2y_2z_2$ is simply

$$\boldsymbol{\omega} = \boldsymbol{\omega}_1 + \boldsymbol{\omega}_2 \,. \tag{2.73}$$

Let us assume that the vector $\boldsymbol{\omega}$ is expressed in terms of components along the frame $x_1y_1z_1$; the angular acceleration then consists of two parts: the first due to the change in the component of $\boldsymbol{\omega}$ relative to the frame $x_1y_1z_1$ and the second due to the fact that $\boldsymbol{\omega}$ is expressed in terms of components along a rotating frame. Denoting the first part by $\boldsymbol{\alpha}'$ and noting that the second part can be obtained from (2.67), by replacing $\boldsymbol{r}$ by $\boldsymbol{\omega}$ and $\boldsymbol{\omega}$ by $\boldsymbol{\omega}_1$, yields

$$\boldsymbol{\alpha} = \dot{\boldsymbol{\omega}} = \boldsymbol{\alpha}' + \boldsymbol{\omega}_1 \times \boldsymbol{\omega} = \boldsymbol{\alpha}' + \boldsymbol{\omega}_1 \times \boldsymbol{\omega}_2 \,. \tag{2.74}$$

Example 2.3: A bicycle travels on a circular track of radius R with the circumferential velocity $\boldsymbol{v}_{\mathrm{A}}$ and acceleration $\boldsymbol{a}_{\mathrm{A}}$. Figure 2.26 shows one of the bicycle wheels rotating on the vertical $\boldsymbol{jk}$ plane. The radius of the wheel is r. Determine the velocity $\boldsymbol{v}$ and acceleration $\boldsymbol{a}$ of a point P on the tire when the radius from the center of the wheel at the point P makes an angle θ with respect to the horizontal plane.

Two reference frames are employed. The frame $x_1y_1z_1$ is attached to the bicycle, and the frame $x_2y_2z_2$ is attached to the wheel. The two frames coincide instantaneously, although the frame $x_2y_2z_2$ rotates relative to the frame $x_1y_1z_1$.

The velocity of the wheel center is

$$\boldsymbol{v}_{\mathrm{A}} = -v_{\mathrm{A}}\boldsymbol{j} \,. \tag{2.75a}$$

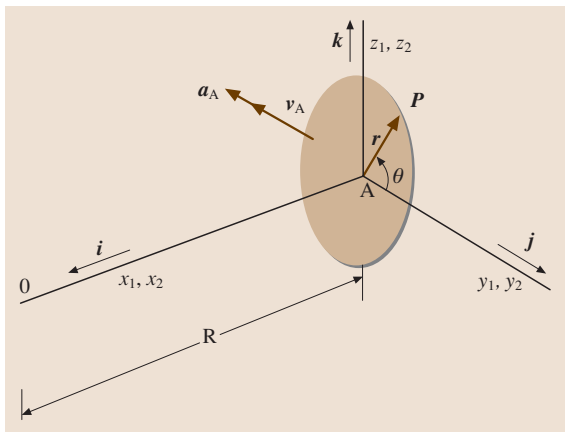


Fig. 2.26 Bicycle wheel

and the angular velocity of the frame $x_1y_1z_1$ is

$$\boldsymbol{\omega}_1 = \frac{v_A}{R}\boldsymbol{k}\,. \tag{2.75b}$$

The angular velocity of the frame $x_2y_2z_2$ relative to $x_1y_1z_1$ is

$$\boldsymbol{\omega}_2 = \frac{v_A}{r}\boldsymbol{i}\,. \tag{2.75c}$$

So, the absolute angular velocity of the frame $x_2y_2z_2$ is

$$\boldsymbol{\omega} = \boldsymbol{\omega}_1 + \boldsymbol{\omega}_2 = \frac{v_A}{R}\boldsymbol{k} + \frac{v_A}{r}\boldsymbol{i}\,. \tag{2.75d}$$

The radius vector from A to P is

$$\boldsymbol{r}_{AP} = r(\cos\theta\boldsymbol{j} + \sin\theta\boldsymbol{k})\,. \tag{2.75e}$$

The velocity of P becomes

$$\begin{aligned}\boldsymbol{v} &= v_A + \boldsymbol{\omega}\times\boldsymbol{r}_{AP}\\ &= -v_A\boldsymbol{j} + \left(\frac{v_A}{r}\boldsymbol{i} + \frac{v_A}{R}\boldsymbol{k}\right)\times r(\cos\theta\boldsymbol{j} + \sin\theta\boldsymbol{k})\\ &= -\frac{v_A r}{R}\cos\theta\boldsymbol{i} - v_A(1+\sin\theta)\boldsymbol{j} + v_A\cos\theta\boldsymbol{k}\,.\end{aligned} \tag{2.75f}$$

The acceleration of A has two components: a tangential component due to the acceleration of the bicycle along the track and a normal component due to motion along a curvilinear track. Therefore,

$$\begin{aligned}a_A &= -a_A\boldsymbol{j} + \boldsymbol{\omega}_1\times(\boldsymbol{\omega}_1\times\boldsymbol{r}_{0A})\\ &= -a_A\boldsymbol{j} + \frac{v_A}{R}\boldsymbol{k}\times\left[\frac{v_A}{R}\boldsymbol{k}\times(-R)\boldsymbol{i}\right]\\ &= \frac{v_A^2}{R}\boldsymbol{i} - a_A\boldsymbol{j}\,.\end{aligned} \tag{2.75g}$$

Using (2.74), the angular acceleration of the frame $x_2y_2z_2$ is

$$\begin{aligned}\boldsymbol{\alpha} &= \dot{\boldsymbol{\omega}}_1 + \dot{\boldsymbol{\omega}}_2 + \boldsymbol{\omega}_1\times\boldsymbol{\omega}_2\\ &= \frac{a_A}{R}\boldsymbol{k} + \frac{a_A}{r}\boldsymbol{i} + \frac{v_A}{R}\boldsymbol{k}\times\frac{v_A}{r}\boldsymbol{i}\\ &= \frac{a_A}{r}\boldsymbol{i} + \frac{v_A^2}{rR}\boldsymbol{j} + \frac{a_A}{R}\boldsymbol{k}\,.\end{aligned} \tag{2.75h}$$

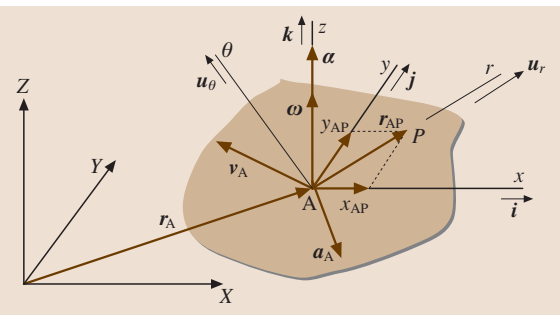


Fig. 2.27 Rigid body motion in a plane

Finally, the acceleration of P is

$$\begin{aligned}\boldsymbol{a} &= \boldsymbol{a}_A + \boldsymbol{\alpha}\times\boldsymbol{r}_{AP} + \boldsymbol{\omega}\times(\boldsymbol{\omega}\times\boldsymbol{r}_{AP})\\ &= \frac{v_A^2}{R}\boldsymbol{i} - a_A\boldsymbol{j} + \left(\frac{a_A}{r}\boldsymbol{i} + \frac{v_A^2}{rR}\boldsymbol{j} + \frac{a_A}{R}\boldsymbol{k}\right)\\ &\quad\times r(\cos\theta\boldsymbol{j} + \sin\theta\boldsymbol{k}) + \left(\frac{v_A}{r}\boldsymbol{i} + \frac{v_A}{R}\boldsymbol{k}\right)\\ &\quad\times\left[\left(\frac{v_A}{r}\boldsymbol{i} + \frac{v_A}{R}\boldsymbol{k}\right)\times r(\cos\theta\boldsymbol{j} + \sin\theta\boldsymbol{k})\right]\\ &= \left[\frac{v_A^2}{R}(1+2\sin\theta) - \frac{ra_A}{R}\cos\theta\right]\boldsymbol{i}\\ &\quad - \left[\left(\frac{1}{r} + \frac{r}{R^2}\right)v_A^2\cos\theta + a_A + a_A\sin\theta\right]\boldsymbol{j}\\ &\quad - \left(\frac{v_A^2}{r}\sin\theta - a_A\cos\theta\right)\boldsymbol{k}\,.\end{aligned} \tag{2.75i}$$

2.2.5 Planar Motion of a Rigid Body

A body has three degrees of motion in planar motion: two of translation (displacement in the x- and y-directions) and one of rotation (rotation about the z-axis) as shown in Fig. 2.27. The radius vector $\boldsymbol{r}_{AP}$ is written as

$$\boldsymbol{r}_{AP} = x_{AP}\boldsymbol{i} + y_{AP}\boldsymbol{j}\,. \tag{2.76}$$

The angular velocity and acceleration of point P are

$$\boldsymbol{\omega} = \omega\boldsymbol{k}\,, \quad \boldsymbol{\alpha} = \alpha k\,, \tag{2.77}$$

respectively. The velocity of point P is

$$\begin{aligned}\boldsymbol{v} &= \boldsymbol{v}_A + \boldsymbol{\omega}\times\boldsymbol{r}_{AP} = \boldsymbol{v}_A + \omega\boldsymbol{k}\times(x_{AP}\boldsymbol{i} + y_{AP}\boldsymbol{j})\\ &= \boldsymbol{v}_A - \omega y_{AP}\boldsymbol{i} + \omega x_{AP}\boldsymbol{j}\,.\end{aligned} \tag{2.78}$$

and the acceleration of point P is

$$\begin{aligned}\boldsymbol{a} &= \boldsymbol{a}_A + \alpha\times\boldsymbol{r}_{AP} + \boldsymbol{\omega}\times(\boldsymbol{\omega}\times\boldsymbol{r}_{AP})\\ &= a_A + \alpha\boldsymbol{k}\times(x_{AP}\boldsymbol{i} + y_{AP}\boldsymbol{j})\\ &\quad + \omega\boldsymbol{k}\times[\omega\boldsymbol{k}\times(x_{AP}\boldsymbol{i} + y_{AP}\boldsymbol{j})]\\ &= \boldsymbol{a}_A - \alpha(y_{AP}\boldsymbol{i} - x_{AP}\boldsymbol{j}) - \omega^2\boldsymbol{r}_{AP}\,.\end{aligned} \tag{2.79}$$

The motion can also be expressed in terms of radial and transverse components. The radius vector is

$$\boldsymbol{r}_{AP} = r_{AP}\boldsymbol{u}_r\,, \tag{2.80}$$

where $\boldsymbol{u}_r$ is the unit vector in the radial direction. The velocity of point P is

$$\boldsymbol{v} = \boldsymbol{v}_A + \omega r_{AP}\boldsymbol{u}_\theta\,, \tag{2.81}$$

where $\boldsymbol{u}_\theta$ is the unit vector in the transverse direction. The acceleration of point P is

$$\begin{aligned}\boldsymbol{a} &= \boldsymbol{a}_\mathrm{A} + \alpha r_\mathrm{AP}\boldsymbol{u}_\theta - \omega^2 \boldsymbol{r}_\mathrm{AP} \\ &= \boldsymbol{a}_\mathrm{A} + \alpha r_\mathrm{AP}\boldsymbol{u}_\theta - \omega^2 r_\mathrm{AP}\boldsymbol{u}_\mathrm{r} \,. \end{aligned} \tag{2.82}$$

Equation (2.78) consists of two terms. The first term represents the velocity of translation of a reference point A. The second term represents the velocity due to rotation about A. There exists a point C such that the velocity of P can be regarded instantaneously as due entirely to rotation about C. It follows that the point C is instantaneously at rest. The point C is the *instantaneously center of rotation*. Point C may lie inside or outside the body. If both the magnitude and direction of the velocity vector are known, and the angular velocity is also given, then the instantaneously center can be determined by a graphical approach. Figure 2.28 depicts the rigid body and the velocity at point P. The velocity vector can be written

$$\boldsymbol{v} = \boldsymbol{\omega}\times\boldsymbol{r} = \omega\boldsymbol{k}\times r\boldsymbol{u}_r = \omega r\boldsymbol{u}_\theta = v\boldsymbol{u}_\theta \,. \tag{2.83}$$

The radius vector $\boldsymbol{r}$ is normal to the velocity vector $\boldsymbol{v}$, and the angular velocity ω is in the counterclockwise direction as shown in Fig. 2.28. The magnitude of the radius vector is

$$r = \frac{v}{\omega} \,. \tag{2.84}$$

The instantaneously center of rotation describes the *space centrode* during motion in relation to a coordinate system fixed in space and in relation to a fixed-body coordinate system of the *body centrode*. During motion, the body centrode rolls on the space centrode.

Example 2.4: A slipping bar of length $L = 20\,\mathrm{m}$ is depicted in Fig. 2.29. The velocity of the point A has magnitude $v_\mathrm{A} = 40\,\mathrm{m/s}$ when the angle between the bar and the wall is $\theta = 30°$; determine the angular velocity of the bar and the velocity of the point B and plot the body and space centrodes.

The instantaneously center of rotation lies at the intersection of the normal to the wall at A and the normal to the floor at the point B. From Fig. 2.29, one has

$$\begin{aligned} r_\mathrm{A} &= L\sin\theta = 10\,\mathrm{m} \,, \\ r_\mathrm{B} &= L\cos\theta = 17.32\,\mathrm{m} \,. \end{aligned} \tag{2.85}$$

The velocity vector $\boldsymbol{v}_\mathrm{A}$ is in the negative y-direction and the angular velocity in the counterclockwise direction is

$$\omega = \frac{v_\mathrm{A}}{r_\mathrm{A}} = 4\,\mathrm{rad/s} \,. \tag{2.86}$$

The velocity vector $\boldsymbol{v}_\mathrm{B}$ is in the x-direction and its magnitude is

$$v_\mathrm{B} = \omega r_\mathrm{B} = 69.28\,\mathrm{m/s} \,. \tag{2.87}$$

The points A, 0, B, and C are the corners of a rectangular with diagonals equal to L, as shown in Fig. 2.29. Point C is always at a distance L from 0. Therefore, the space centrode is one quarter of a circle with radius L and the center at 0, as depicted by the solid line in Fig. 2.29. At the same time, the velocity vectors $\boldsymbol{v}_\mathrm{A}$ and $\boldsymbol{v}_\mathrm{B}$ make a 90° angle at C. The body centrode is the

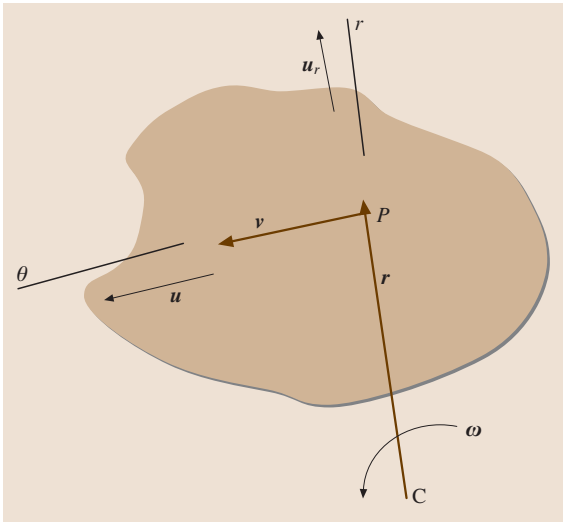


Fig. 2.28 Rigid body and the velocity vector

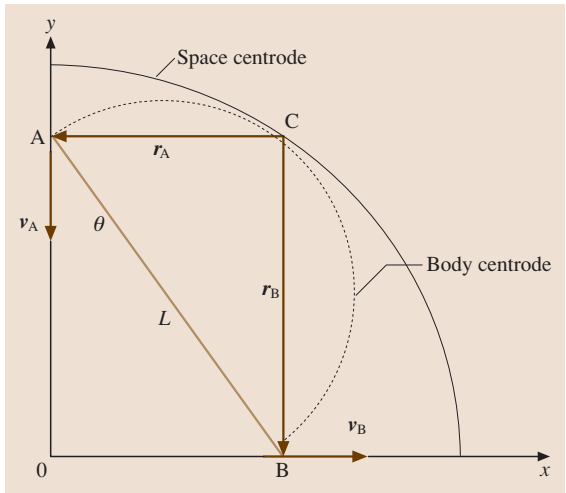


Fig. 2.29 A sliding rod

locus of the point C that is a semicircle as depicted by the dashed line in Fig. 2.29. As the bar slides, the body centrode rolls on the space centrode.

2.2.6 General Case of Motion

Consider the case that the particle P is no longer at rest relative to the moving frame xyz, but can move relative to that frame as depicted in Fig. 2.30. Given an arbitrary vector $\boldsymbol{r}$

$$\boldsymbol{r} = x\boldsymbol{i} + y\boldsymbol{j} + z\boldsymbol{k} \,, \tag{2.88}$$

where x, y, and z are the Cartesian components of the vector and $\boldsymbol{i}$, $\boldsymbol{j}$, and $\boldsymbol{k}$ are the unit vectors along these axes. The unit vectors $\boldsymbol{i}$, $\boldsymbol{j}$, and $\boldsymbol{k}$ rotate with the same angular velocity ω as the moving frame. Hence, from (2.67), we have

$$\begin{aligned}\dot{\boldsymbol{r}} &= \dot{x}\boldsymbol{i} + \dot{y}\boldsymbol{j} + \dot{z}\boldsymbol{k} + x\dot{\boldsymbol{i}} + y\dot{\boldsymbol{j}} + z\dot{\boldsymbol{k}} \\ &= \dot{x}\boldsymbol{i} + \dot{y}\boldsymbol{j} + \dot{z}\boldsymbol{k} + \boldsymbol{\omega}\times(x\boldsymbol{i} + y\boldsymbol{j} + z\boldsymbol{k}) \\ &= \dot{\boldsymbol{r}}' + \boldsymbol{\omega}\times\boldsymbol{r} \,,\end{aligned} \tag{2.89}$$

where

$$\dot{\boldsymbol{r}}' = \dot{x}\boldsymbol{i} + \dot{y}\boldsymbol{j} + \dot{z}\boldsymbol{k} \tag{2.90}$$

is the time rate of change of $\boldsymbol{r}$ regarding the reference frame xyz as inertial.

The position vector of point P as depicted in Fig. 2.30 is

$$\boldsymbol{R} = \boldsymbol{r}_\mathrm{A} + \boldsymbol{r}_\mathrm{AP} \,. \tag{2.91}$$

The absolute velocity of P is

$$\boldsymbol{v} = \dot{\boldsymbol{R}} = \dot{\boldsymbol{r}}_\mathrm{A} + \dot{\boldsymbol{r}}_\mathrm{AP} = \boldsymbol{v}_\mathrm{A} + \boldsymbol{v}'_\mathrm{AP} + \boldsymbol{\omega}\times\boldsymbol{r}_\mathrm{AP} \,, \tag{2.92}$$

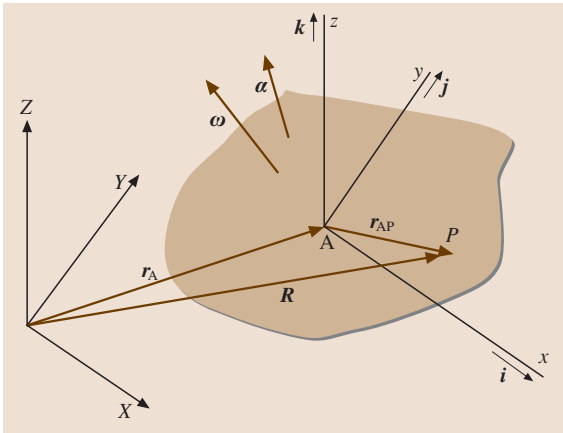


Fig. 2.30 General motion

where $\boldsymbol{v}_\mathrm{A}$ is the velocity of point A relative to the inertial space, and

$$\boldsymbol{v}'_\mathrm{AP} = \dot{x}_\mathrm{AP}\boldsymbol{i} + \dot{y}_\mathrm{AP}\boldsymbol{j} + \dot{z}_\mathrm{AP}\boldsymbol{k} \tag{2.93}$$

is the velocity of P relative to the moving frame xyz, where x_AP, y_AP, and z_AP are the Cartesian components of $\boldsymbol{r}_\mathrm{AP}$, and $\omega\times\boldsymbol{r}_\mathrm{AP}$ is the velocity of P entirely due to the rotation of the frame xyz. Similarly, the absolute acceleration of P is

$$\begin{aligned}\boldsymbol{a} = \dot{\boldsymbol{v}} &= \dot{\boldsymbol{v}}_\mathrm{A} + \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{v}'_\mathrm{AP} + \dot{\boldsymbol{\omega}}\times\boldsymbol{r}_\mathrm{AP} + \boldsymbol{\omega}\times\dot{\boldsymbol{r}}_\mathrm{AP} \\ &= \boldsymbol{a}_\mathrm{A} + \boldsymbol{a}'_\mathrm{AP} + \boldsymbol{\omega}\times\boldsymbol{v}'_\mathrm{AP} + \boldsymbol{\alpha}\times\boldsymbol{r}_\mathrm{AP} + \boldsymbol{\omega}\times\boldsymbol{v}'_\mathrm{AP} \\ &\quad + \boldsymbol{\omega}\times(\boldsymbol{\omega}\times\boldsymbol{r}_\mathrm{AP}) \\ &= a_\mathrm{A} + a'_\mathrm{AP} + 2\boldsymbol{\omega}\times\boldsymbol{v}'_\mathrm{AP} + \boldsymbol{\alpha}\times\boldsymbol{r}_\mathrm{AP} \\ &\quad + \boldsymbol{\omega}\times(\boldsymbol{\omega}\times\boldsymbol{r}_\mathrm{AP}) \,,\end{aligned} \tag{2.94}$$

where

$$\boldsymbol{a}'_\mathrm{AP} = \ddot{x}_\mathrm{AP}\boldsymbol{i} + \ddot{y}_\mathrm{AP}\boldsymbol{j} + \ddot{z}_\mathrm{AP}\boldsymbol{k} \tag{2.95}$$

is the acceleration of P relative to the moving frame xyz, $2\boldsymbol{\omega}\times\boldsymbol{v}'_\mathrm{AP}$ is the Coriolis acceleration, and $\boldsymbol{\alpha}\times\boldsymbol{r}_\mathrm{AP} + \boldsymbol{\omega}\times(\boldsymbol{\omega}\times\boldsymbol{r}_\mathrm{AP})$ is the acceleration of P entirely due to the rotation of the frame xyz, where $\boldsymbol{\alpha} = \dot{\boldsymbol{\omega}}$ is the angular acceleration of the frame xyz.

2.2.7 Dynamics

Dynamics of a Particle

Particle Dynamics. Dynamics describes the motion of mass particles, mass particle systems, bodies and body systems, in terms of the forces and moments, under the laws of kinematics [2.5–9].

Newton's Law of Motion. Newton's law of motion can be applied to systems of particles and rigid bodies. Newton suggested the concept of inertial systems of reference, i. e., systems of reference that are either at rest or moving with uniform velocity relative to a fixed reference frame. The motion of any particle is measured relative to such an inertial system and is said to be *absolute*. The linear momentum vector $\boldsymbol{p}$ is defined as the product of the mass m of the particle and the absolute velocity $\boldsymbol{v}$, or $\boldsymbol{p} = m\boldsymbol{v}$. The second law is

$$\boldsymbol{F} = \frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}m\boldsymbol{v} \,. \tag{2.96}$$

In SI units, the unit of mass is the kilogram (kg) and the unit of force is Newton (N). If the mass m is constant, then

$$\boldsymbol{F} = m\frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t} = m\boldsymbol{a} \,, \tag{2.97}$$

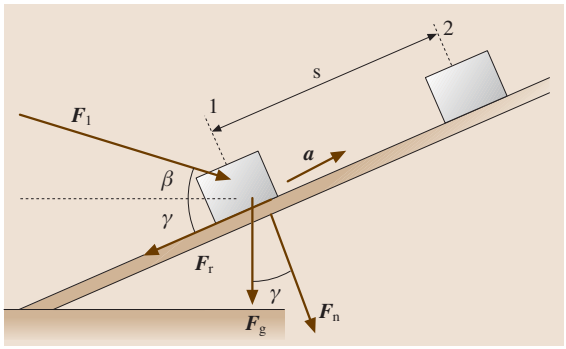


Fig. 2.31 Point mass on an inclined plane

where $\boldsymbol{a} = \mathrm{d}\boldsymbol{v}/\mathrm{d}t$ is the absolute acceleration of m. Equation (2.97) is the equation of motion of a particle.

Example 2.5: The mass $m = 5\,\mathrm{kg}$ is moved from the position of rest 1 by the force $F_1 = 100\,\mathrm{N}$ with $\beta = 15°$ onto the inclined plane with $\gamma = 25°$ as depicted in Fig. 2.31. The friction coefficient is $\mu = 0.3$. Determine the acceleration, velocity, and time upon arrival at position 2 after traveling $s = 8\,\mathrm{m}$.

$$\begin{aligned} \boldsymbol{F}_n &= m\boldsymbol{g}\cos\gamma + \boldsymbol{F}_1 \sin(\beta+\gamma) = 108.7\,\mathrm{N} \\ m\boldsymbol{a} &= \sum \boldsymbol{F} = \boldsymbol{F}_1\cos(\beta+\gamma) - \boldsymbol{F}_g \sin\gamma - \mu \boldsymbol{F}_n \\ &= 23.3\,\mathrm{N}\,. \end{aligned} \tag{2.98a}$$

In scalar notation, we have

$$\begin{aligned} a &= \frac{ma}{m} = 4.66\,\mathrm{m/s^2} \quad t = \sqrt{2s/a} = 1.85\,\mathrm{s}\,, \\ v &= \sqrt{2as} = 8.63\,\mathrm{m/s}\,. \end{aligned} \tag{2.98b}$$

Basic Concepts of Energy, Work, and Power.

Work. From Fig. 2.32, the increment of work $\mathrm{d}W$, a scalar, is defined as the dot product (scalar product) of the force vector and the increment of distance vector

$$\mathrm{d}W = \boldsymbol{F}\cdot\mathrm{d}\boldsymbol{r} = F\cos\beta\,\mathrm{d}r = F_1\,\mathrm{d}r\,. \tag{2.99}$$

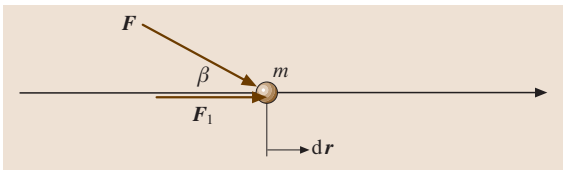


Fig. 2.32 Basic concept of work

From Newton's second law, $\boldsymbol{F} = m\ddot{\boldsymbol{r}}$, and $\mathrm{d}\boldsymbol{r} = \dot{\boldsymbol{r}}\,\mathrm{d}t$, we obtain

$$\begin{aligned} \mathrm{d}W &= m\ddot{\boldsymbol{r}}\cdot\dot{\boldsymbol{r}}\,\mathrm{d}t = m\frac{\mathrm{d}\dot{\boldsymbol{r}}}{\mathrm{d}t}\cdot\dot{\boldsymbol{r}}\,\mathrm{d}t \\ &= m\dot{\boldsymbol{r}}\cdot d\dot{\boldsymbol{r}} = \mathrm{d}\left(\frac{1}{2}m\dot{\boldsymbol{r}}\cdot\dot{\boldsymbol{r}}\right)\,. \end{aligned} \tag{2.100}$$

The kinetic energy T, a scalar, is defined as

$$T = \frac{1}{2}m\dot{\boldsymbol{r}}\cdot\dot{\boldsymbol{r}}\,. \tag{2.101}$$

Consider the work performed by force $\boldsymbol{F}$ in moving the particle m from position $\boldsymbol{r}_1$ to position $\boldsymbol{r}_2$ along curve S as depicted in Fig. 2.33, one has

$$W_{1-2} = \int_{r_1}^{r_2} \boldsymbol{F}\cdot\mathrm{d}\boldsymbol{r} = \int_{T_1}^{T_2} \mathrm{d}T = T_2 - T_1\,. \tag{2.102}$$

If forces have a potential as below

$$\boldsymbol{F} = -\mathbf{grad}\,U = -\frac{\partial U}{\partial x}\boldsymbol{i} - \frac{\partial U}{\partial y}\boldsymbol{j} - \frac{\partial U}{\partial z}\boldsymbol{k}\,, \tag{2.103}$$

then it follows that

$$\begin{aligned} W &= -\int_{P_1}^{P_2}\left(\frac{\partial U}{\partial x}\mathrm{d}x + \frac{\partial U}{\partial y}dy + \frac{\partial U}{\partial z}\mathrm{d}z\right) \\ &= -\int_{P_1}^{P_2}\mathrm{d}U = U_1 - U_2\,. \end{aligned} \tag{2.104}$$

In this case, work is independent of the integration distance and equal to the difference of the potential between the initial point P_1 and the final point P_2, as depicted in Fig. 2.34a. Forces with potential are forces of gravity and spring forces.

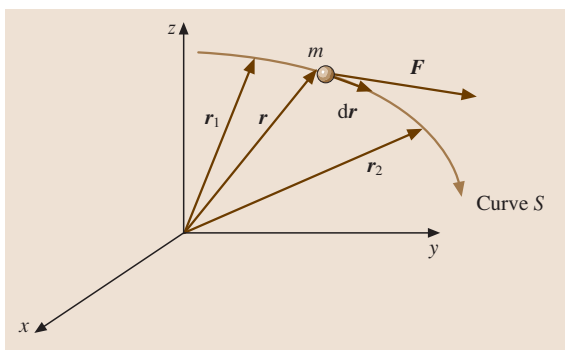


Fig. 2.33 Work of a force

Special Work Examples.

1. Force of gravity. The potential energy $U = F_g z$ and the work is

$$W_g = -\int_{U_1}^{U_2} dU = U_1 - U_2 = F_g(z_1 - z_2) \,. \tag{2.105}$$

2. Spring force (Fig. 2.34b). Potential spring energy $U = cs^2/2$ with spring constant c. The spring force is $\boldsymbol{F}_c = -\nabla U = -\partial U/\partial s\boldsymbol{i} = -cs\boldsymbol{i}$ and the work is

$$W_c = \int_{s_1}^{s_2} cs\, ds = \frac{c\left(s_2^2 - s_1^2\right)}{2} \,. \tag{2.106}$$

3. Frictional force (Fig. 2.34c). There is no potential since frictional work is lost in the form of heat

$$W_r = \int_{s_1}^{s_2} \boldsymbol{F}_r \cdot d\boldsymbol{r} = \int_{s_1}^{s_2} F_r \cos\pi\, ds = -\int_{s_1}^{s_2} F_r\, ds \,. \tag{2.107}$$

4. Torque (Fig. 2.34d). Only the moment components M_t parallel to the axis of rotation perform work

$$W_M = \int_{\phi_1}^{\phi_2} M(\phi) \cdot d\phi = \int_{\phi_1}^{\phi_2} M(\phi) \cos\gamma\, d\phi$$
$$= \int_{\phi_1}^{\phi_2} M_t(\phi)\, d\phi \,. \tag{2.108}$$

Total Work. If forces and moments are at work on a body simultaneously, then

$$W = \int_{s_1}^{s_2} \sum \boldsymbol{F}_i\, d\boldsymbol{r}_i + \int_{\phi_1}^{\phi_2} \sum \boldsymbol{M}_i\, d\phi_i \,. \tag{2.109}$$

Power. Power is defined as work per unit time

$$P(t) = \frac{dW}{dt} = \sum \boldsymbol{F}_i \boldsymbol{v}_i + \sum \boldsymbol{M}_i \boldsymbol{\omega}_i \,. \tag{2.110}$$

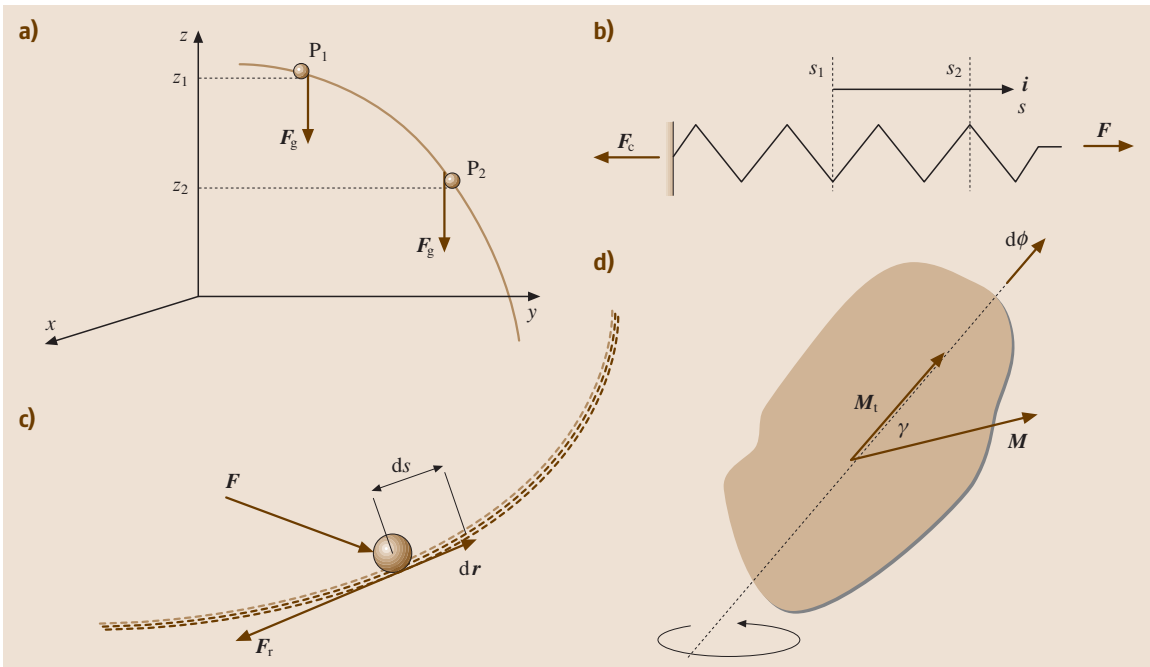


Fig. 2.34a–d Examples of work: (a) gravity, (b) spring force, (c) friction, (d) torque

Mean Power.

$$P_{\mathrm{m}} = \frac{\int_{t_1}^{t_2} P(t)\,\mathrm{d}t}{t_2 - t_1} = \frac{W}{t_2 - t_1} \,. \tag{2.111}$$

2.2.8 Straight-Line Motion of Particles and Rigid Bodies

Momentum Equation

From (2.96), for constant mass, we have

$$\boldsymbol{P}_{1,2} = \int_{t_1}^{t_2} \boldsymbol{F}\,\mathrm{d}t = \int_{v_1}^{v_2} m\,\mathrm{d}\boldsymbol{v} = m\boldsymbol{v}_2 - m\boldsymbol{v}_1 = \boldsymbol{p}_2 - \boldsymbol{p}_1 \,. \tag{2.112}$$

The time integral of the force, known as the *linear impulse vector*, is equal to the difference in momentum.

Angular Momentum Equation

Consider a particle of mass m moving under the action of a force $\boldsymbol{F}$. From Fig. 2.35, it shows the position of m relative to the origin 0 of the inertial frame xyz by $\boldsymbol{r}$ and the absolute velocity of m by $\boldsymbol{v}$. The moment of momentum or angular momentum of m with respect to point 0 is defined as the moment of the linear momentum $\boldsymbol{p}$ about 0 and is represented by the cross product of the vectors $\boldsymbol{r}$ and $\boldsymbol{p}$. The angular momentum of m about point 0 is

$$\boldsymbol{H}_0 = \boldsymbol{r} \times \boldsymbol{p} = \boldsymbol{r} \times m\boldsymbol{v} = \boldsymbol{r} \times m\dot{\boldsymbol{r}} \,. \tag{2.113}$$

Assuming that m is a constant, we have

$$\dot{\boldsymbol{H}}_0 = \dot{\boldsymbol{r}} \times m\dot{\boldsymbol{r}} + \boldsymbol{r} \times m\ddot{\boldsymbol{r}} = \boldsymbol{r} \times m\ddot{\boldsymbol{r}} = \boldsymbol{r} \times \boldsymbol{F} = \boldsymbol{M}_0 \,. \tag{2.114}$$

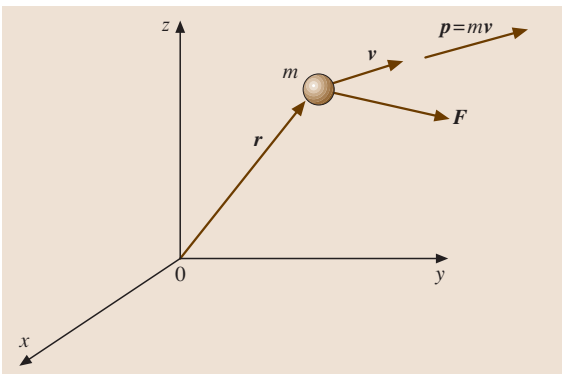


Fig. 2.35 Angular momentum of a particle

Note that $\dot{\boldsymbol{r}} \times m\dot{\boldsymbol{r}} = m(\dot{\boldsymbol{r}} \times \dot{\boldsymbol{r}}) = 0$ and $\boldsymbol{r} \times \boldsymbol{F} = \boldsymbol{M}_0$ is the moment of the force about 0. Therefore, the moment of a force about 0 is equal to the time rate of change of the moment of momentum about 0.

The angular impulse vector about 0 between the times t_1 and t_2 is

$$\hat{\boldsymbol{M}}_0 = \int_{t_1}^{t_2} \boldsymbol{M}_0\,\mathrm{d}t = \int_{t_1}^{t_2} \frac{\mathrm{d}\boldsymbol{H}_0}{\mathrm{d}t}\,\mathrm{d}t = \boldsymbol{H}_0(t_2) - \boldsymbol{H}_0(t_1) = \Delta\boldsymbol{H}_0 \,. \tag{2.115}$$

Therefore, the angular impulse vector about 0 between the times t_1 and t_2 is equal to the change in the angular momentum vector about 0 between the same two instants.

2.2.9 Dynamics of Systems of Particles

A system of particles is a group of n particles as shown in Fig. 2.36. The external and internal forces are denoted by $\boldsymbol{F}_i$ and $\boldsymbol{f}_i$, respectively. The internal force is the resultant of the interaction forces $\boldsymbol{f}_{ij}$ exerted by the particles $m_j (j = 1, 2, \ldots, n,\ j \neq i)$ on particle $m_i (i = 1, 2, \ldots, n)$. The equation of motion of the system of particles is

$$\sum_{i=1}^{n} \boldsymbol{F}_i + \sum_{j=1}^{n}\sum_{i=1}^{n} \boldsymbol{f}_{ij} = \sum_{i=1}^{n} \boldsymbol{F}_i = \sum_{i=1}^{n} m_i \ddot{\boldsymbol{r}}_i = \sum_{i=1}^{n} m_i \boldsymbol{a}_i \,. \tag{2.116}$$

By Newton's third law, $\boldsymbol{f}_{ij} = -\boldsymbol{f}_{ji}$. Hence $\sum_{j=1}^{n} \sum_{i=1}^{n} \boldsymbol{f}_{ij} = 0$.

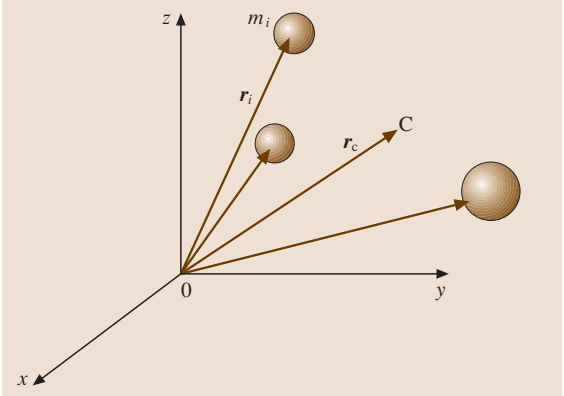


Fig. 2.36 A group of particles

The center of mass C of the system is a point in space representing a weighted average position of the system. The weighting factor is the mass of the particle. We have

$$\boldsymbol{F} = m\boldsymbol{a}_\mathrm{c} \,, \tag{2.117}$$

where

$$\boldsymbol{F} = \sum_{i=1}^{n} \boldsymbol{F}_i m = \sum_{i=1}^{n} m_i \boldsymbol{a}_\mathrm{c} = \frac{1}{m} \sum_{i=1}^{n} m_i \ddot{\boldsymbol{r}}_i \,. \tag{2.118}$$

2.2.10 Momentum Equation

From Fig. 2.36, the linear momentum of m_i is

$$\boldsymbol{p}_i = m_i \boldsymbol{v}_i \,. \tag{2.119}$$

The linear momentum of the system is

$$\boldsymbol{p} = \sum_{i=1}^{n} \boldsymbol{p}_i = \sum_{i=1}^{n} m_i \boldsymbol{v}_i = m\boldsymbol{v}_\mathrm{c} \,. \tag{2.120}$$

The resultant of the external forces acting on the system is

$$\boldsymbol{F} = \dot{\boldsymbol{p}} = m\dot{\boldsymbol{v}}_\mathrm{c} = m\boldsymbol{a}_\mathrm{c} \,. \tag{2.121}$$

If $\boldsymbol{F} = 0$, then $\boldsymbol{p} =$ constant. This is the conservation of linear momentum of a system of particles.

The angular momentum of the particle m_i about 0 is

$$\boldsymbol{H}_{0i} = \boldsymbol{r}_i \times \boldsymbol{p}_i = \boldsymbol{r}_i \times m_i \boldsymbol{v}_i \,. \tag{2.122}$$

The angular momentum of the system about 0 is

$$\boldsymbol{H}_0 = \sum_{i=1}^{n} \boldsymbol{H}_{0i} = \sum_{i=1}^{n} \boldsymbol{r}_i \times m_i \boldsymbol{v}_i \,. \tag{2.123}$$

Hence

$$\begin{aligned} \dot{\boldsymbol{H}}_0 &= \sum_{i=1}^{n} \dot{\boldsymbol{r}}_i \times m_i \boldsymbol{v}_i + \sum_{i=1}^{n} \boldsymbol{r}_i \times m_i \dot{\boldsymbol{v}}_i \\ &= \sum_{i=1}^{n} \boldsymbol{r}_i \times m_i \boldsymbol{a}_i = \sum_{i=1}^{n} \boldsymbol{r}_i \times \boldsymbol{F}_i \,. \end{aligned} \tag{2.124}$$

Since $\boldsymbol{M}_0 = \sum_{i=1}^{n} \boldsymbol{r}_i \times \boldsymbol{F}_i$, we have

$$\boldsymbol{M}_0 = \dot{\boldsymbol{H}}_0 \,. \tag{2.125}$$

If $\boldsymbol{M}_0 = 0$, then $\boldsymbol{H}_0 =$ constant. This states that, in the absence of external torques about 0, the angular momentum of the system about 0 is a constant. This is the conservation of angular momentum of the system about a fixed point. It can be extended to the conservation of angular momentum about the mass center.

Energy Equation

The kinetic energy of particle m_i is

$$T_i = \frac{1}{2} m_i \dot{\boldsymbol{r}}_i \cdot \dot{\boldsymbol{r}}_i \,. \tag{2.126}$$

and the kinetic energy of the system is

$$T = \sum_{i=1}^{n} T_i = \frac{1}{2} \sum_{i=1}^{n} m_i \dot{\boldsymbol{r}}_i \cdot \dot{\boldsymbol{r}}_i \,. \tag{2.127}$$

Example 2.6: A spring with spring constant c, which is pre-stressed by the value s, thrusts the masses m_1 and m_2 apart from rest as depicted in Fig. 2.37. Disregarding the friction forces during the relaxation process of the spring, there is no external force. Determine the velocities of m_1 and m_2.

From the conservation of momentum of a system of particles, we have $m_1 v_1 - m_2 v_2 = 0$.

The energy equation is

$$\frac{1}{2} c s^2 = \frac{1}{2} \left(m_1 v_1^2 + m_2 v_2^2 \right) . \tag{2.128a}$$

It follows that

$$v_1 = \sqrt{\frac{c s^2}{(m_1 + m_1^2/m_2)}} \,, \quad v_2 = \sqrt{\frac{c s^2}{(m_2 + m_2^2/m_1)}} \,. \tag{2.128b}$$

Example 2.7: Two masses, connected by an inextensible chain, are drawn out of the position of rest by the force $\boldsymbol{F}$ as depicted in Fig. 2.38. Mass m_1 moves along the inclined surface. Determine the velocity after traveling a distance s_1.

The friction force on masses m_1 and m_2 are $F_{\mathrm{r}1} = \mu_1 (F_{\mathrm{g}1} \cos\gamma_1 - F \sin\gamma_1)$ and $F_{\mathrm{r}2} = \mu_2 F_{\mathrm{g}2} \cos\gamma_2$, respectively. As a precondition for the mass m_1 not being lifted, we must have $F \le F_{\mathrm{g}1} \cot\gamma_1$.

The energy equation is

$$\begin{aligned} &F \cos\gamma_1 s_1 + F_{\mathrm{g}1} h_1 - F_{\mathrm{r}1} s_1 - F_{\mathrm{g}2} h_2 - F_{\mathrm{r}2} s_2 \\ &= \frac{1}{2} \left(m_1 v_1^2 + m_2 v_2^2 \right) . \end{aligned} \tag{2.129a}$$

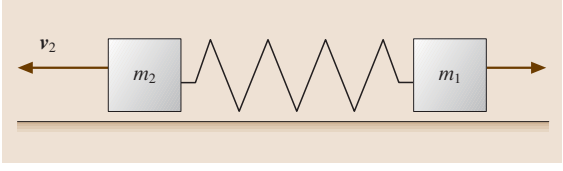


Fig. 2.37 Spring–mass system

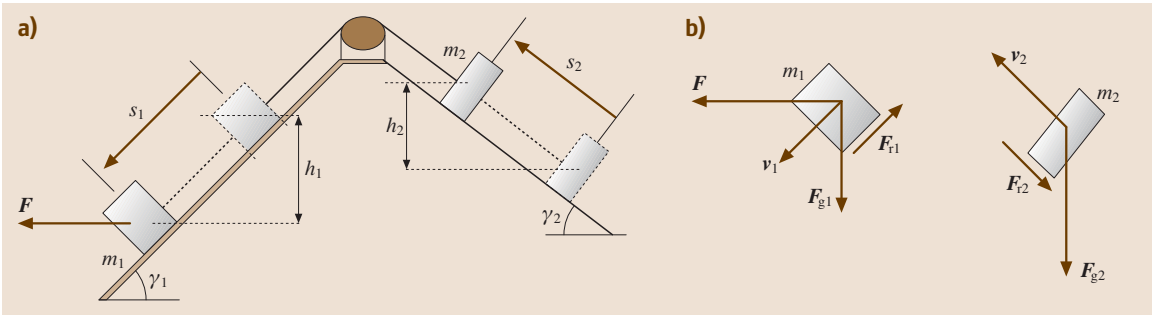


Fig. 2.38 (a) Two-mass system, (b) forces on each mass

Because $s_1 = s_2$, $v_1 = v_2$, $h_1 = s_1 \sin\gamma_1$, and $h_2 = s_2 \sin\gamma_2$, it follows that

$$v_1 = \sqrt{\frac{2s_1(F\cos\gamma_1 + F_{g1}\sin\gamma_1 - F_{r1} - F_{g2}\sin\gamma_2 - F_{r2})}{m_1 + m_2}} \,. \qquad (2.129b)$$

2.2.11 D'Alembert's Principle, Constrained Motion

From Newton's law we know that $\boldsymbol{F} - m\boldsymbol{a} = 0$, i. e., the external forces and forces of inertia of a particle form a *state of equilibrium*. In the event of a system of particles m_i $(i = 1, 2, \ldots, n)$, Newton's equations of motion are

$$\boldsymbol{F}_i + \boldsymbol{f}_i - m_i\ddot{\boldsymbol{r}}_i = 0 \,, \quad i = 1, 2, \ldots, n \,, \qquad (2.130)$$

where $\boldsymbol{F}_i$ are the applied forces, $\boldsymbol{f}_i$ are the constraint forces, and $-m_i\ddot{\boldsymbol{r}}_i$ are the inertial forces. Equation (2.130) is the dynamic equilibrium of the system of particles. The sum of virtual work for the entire system is

$$\sum_{i=1}^{n} (\boldsymbol{F}_i + \boldsymbol{f}_i - m_i\ddot{\boldsymbol{r}}_i)\cdot\delta\boldsymbol{r}_i = 0 \,. \qquad (2.131)$$

However, the virtual work performed by the constraint forces over virtual displacements is zero. Hence, it follows that

$$\sum_{i=1}^{n} (\boldsymbol{F}_i - m_i\ddot{\boldsymbol{r}}_i)\cdot\delta\boldsymbol{r}_i = 0 \,. \qquad (2.132a)$$

Equation (2.132a) is D'Alembert's principle for a system of particles. It can also be applied to a system of rigid bodies. If the motion is planar, the D'Alembert's principle for a system of rigid bodies is

$$\sum_{i=1}^{n} (\boldsymbol{F}_i - m_i\ddot{\boldsymbol{r}}_{Ci})\cdot\delta\boldsymbol{r}_{Ci} + (M_{Ci} - I_{Ci}\ddot{\theta}_i)\delta\theta_i = 0 \,, \qquad (2.132b)$$

where r_{Ci} is the position of the mass center of the i-th rigid body, M_{Ci} is the moment of the mass center of the i-th rigid body, and I_{Ci} is the mass moment of inertia about an axis normal to the plane of motion that passes through C.

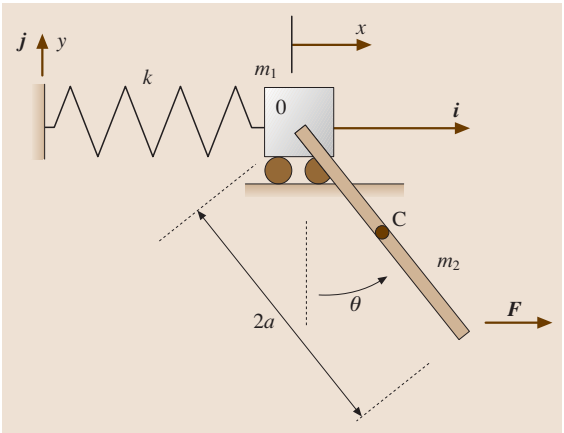


Fig. 2.39 Spring–mass and rod system

Example 2.8: Derive the equation of motion of the system shown in Fig. 2.39 by using D'Alembert's principle. Use θ and x as independent coordinates.

There are two rigid bodies in Fig. 2.39. m_1 can be considered as a particle that is subjected to no moments and no moment of inertia. m_2 is a rigid body. Equation (2.131) becomes

$$(\boldsymbol{F}_1 - m_1\ddot{\boldsymbol{r}}_{C1})\cdot\delta\boldsymbol{r}_{C1} + (\boldsymbol{F}_2 - m_2\ddot{\boldsymbol{r}}_{C2})\cdot\delta\boldsymbol{r}_{C2} + (M_{C2} - I_{C2}\ddot{\theta}_2)\delta\theta_2 = 0 \,. \qquad (2.133a)$$

From Fig. 2.39, we have

$$\boldsymbol{F}_1 = -kx\boldsymbol{i} - m_1 g\boldsymbol{j}\ , \quad \boldsymbol{F}_2 = F\boldsymbol{i} - m_2 g\boldsymbol{j}\ ,$$
$$M_{C2} = Fa\cos\theta\ , \quad I_{C2} = \frac{1}{3}m_2 a^2 \qquad (2.133b)$$

and

$$\boldsymbol{r}_{C1} = (L+x)\boldsymbol{i}\ ,$$
$$\boldsymbol{r}_{C2} = (L+x+a\sin\theta)\boldsymbol{i} - a\cos\theta\boldsymbol{j}\ ,$$
$$\delta\boldsymbol{r}_{C1} = \delta x\boldsymbol{i}$$
$$\delta\boldsymbol{r}_{C2} = (\delta x + a\cos\theta\delta\theta)\boldsymbol{i} + a\sin\theta\delta\theta\boldsymbol{j}\ , \qquad (2.133c)$$

where $\theta = \theta_2 \delta\theta = \delta\theta_2$, and L is the fixed length of the spring before moving. The accelerations are

$$\ddot{\boldsymbol{r}}_{C1} = \ddot{x}\boldsymbol{i}\ ,$$
$$\ddot{\boldsymbol{r}}_{C2} = (\ddot{x} + a\ddot{\theta}\cos\theta - a\dot{\theta}^2\sin\theta)\boldsymbol{i}$$
$$+ a(\ddot{\theta}\sin\theta + \dot{\theta}^2\cos\theta)\boldsymbol{j}\ . \qquad (2.133d)$$

Substituting (2.133b)–(2.133d) into (2.133a) and setting each of the coefficients of $\delta\theta$ and δx equal to zero, the equations of motion are

$$(m_1+m_2)\ddot{x} + m_2 a(\ddot{\theta}\cos\theta - \dot{\theta}^2\sin\theta) + kx = F\ ,$$
$$m_2\cos\theta\ddot{x} + \frac{4}{3}m_2 a\ddot{\theta} + m_2 g\sin\theta = 2F\cos\theta\ . \qquad (2.133e)$$

2.2.12 Lagrange's Equations

Lagrange provided the equations of motion for a system by a differentiation process related to the dynamic (kinetic and potential) energy. Considering an n-degree-of-freedom system, Lagrange's equations read

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{q}_i}\right) - \frac{\partial L}{\partial q_i} = Q_i \quad i = 1, 2, \ldots, n\ , \qquad (2.134)$$

where the Lagrangian $L = T - V$, in which T is the kinetic energy, V is the potential energy, q_i are the generalized coordinates of the system, and Q_i are the generalized forces. The Lagrangian approach is very efficient for deriving the equations of motion for both linear and nonlinear systems.

Example 2.9: Derive the equation of motion of the system of Fig. 2.39 using Lagrange's equations.

Suppose $q_1 = x$ and $q_2 = \theta$. First, we need to calculate the virtual work with the nonconservative force $\boldsymbol{F}$. Denoting the point of application of the force by B, the position vector of B is

$$\boldsymbol{r}_B = (L + x + 2a\sin\theta)\boldsymbol{i} - 2a\cos\theta\boldsymbol{j}\ . \qquad (2.135a)$$

The virtual displacement of B is

$$\delta\boldsymbol{r}_B = (\delta x + 2a\cos\theta\delta\theta)\boldsymbol{i} + 2a\sin\theta\delta\theta\boldsymbol{j}\ . \qquad (2.135b)$$

The force $\boldsymbol{F}$ is

$$\boldsymbol{F} = F\boldsymbol{i}\ . \qquad (2.135c)$$

The nonconservative virtual work is

$$\boldsymbol{F}\cdot\delta\boldsymbol{r}_B = F(\delta x + 2a\cos\theta\delta\theta)\ . \qquad (2.135d)$$

The coefficients of $\delta\theta$ and δx are the nonconservative generalized forces, or

$$X = F\ , \quad \Theta = 2Fa\cos\theta\ . \qquad (2.135e)$$

The kinetic energy is

$$T = \frac{1}{2}m_1\boldsymbol{v}_1\cdot\boldsymbol{v}_1 + \frac{1}{2}m_2\boldsymbol{v}_2\cdot\boldsymbol{v}_2 + \frac{1}{2}I_C\dot{\theta}^2$$
$$= \frac{1}{2}m_1\dot{x}^2 + \frac{1}{2}m_2\left[(\dot{x} + a\dot{\theta}\cos\theta)^2\right.$$
$$\left. + (a\dot{\theta}\sin\theta)^2\right] + \frac{1}{2}\frac{1}{12}m_2(2a)^2\dot{\theta}^2$$
$$= \frac{1}{2}(m_1+m_2)\dot{x}^2 + m_2 a\dot{x}\dot{\theta}\cos\theta + \frac{2}{3}m_2 a^2\dot{\theta}^2\ . \qquad (2.135f)$$

The potential energy is

$$V = \frac{1}{2}kx^2 + m_2 ga(1-\cos\theta)\ . \qquad (2.135g)$$

Hence,

$$L = T - V$$
$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{x}}\right) = (m_1+m_2)\ddot{x} + m_2 a(\ddot{\theta}\cos\theta - \dot{\theta}^2\sin\theta)$$
$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{\theta}}\right) = m_2 a(\ddot{x}\cos\theta - \dot{x}\dot{\theta}\sin\theta) + \frac{4}{3}m_2 a^2\ddot{\theta}$$
$$\frac{\partial L}{\partial x} = -kx\ , \quad \frac{\partial L}{\partial \theta} = -m_2 a\sin\theta(\dot{x}\dot{\theta} + g)\ . \qquad (2.135h)$$

From (2.134), the equations of motion are

$$(m_1+m_2)\ddot{x} + m_2 a(\ddot{\theta}\cos\theta - \dot{\theta}^2\sin\theta) + kx = F\ ,$$
$$m_2\ddot{x}\cos\theta + \frac{4}{3}m_2 a\ddot{\theta} + m_2 g\sin\theta = 2F\cos\theta\ . \qquad (2.135i)$$

2.2.13 Dynamics of Rigid Bodies

Rigid bodies can be viewed as a special type of systems of particles, where the distances between any two

particles are rigidly constrained to be constant. The velocity of a point in the rigid body relative to another is due only to the angular velocity of the rigid body.

Linear and Angular Momentum

The angular momentum of a rigid body rotating with the angular velocity $\boldsymbol{\omega}$ about the fixed point 0 is defined as

$$H_0 = \int_m \boldsymbol{r} \times \boldsymbol{v}\, \mathrm{d}m \,. \tag{2.136}$$

From Fig. 2.40, because the velocity $\boldsymbol{v}$ of any point in the body is due entirely to the rotation about 0, (2.136) becomes

$$H_0 = \int_m \boldsymbol{r} \times (\boldsymbol{\omega} \times \boldsymbol{r})\, \mathrm{d}m = \int_m (\boldsymbol{r} \cdot \boldsymbol{r})\boldsymbol{\omega} - (\boldsymbol{r} \cdot \boldsymbol{\omega})\boldsymbol{r}\, \mathrm{d}m \,. \tag{2.137}$$

Note that $\boldsymbol{A} \times (\boldsymbol{B} \times \boldsymbol{C}) = (\boldsymbol{A} \cdot \boldsymbol{C})\boldsymbol{B} - (\boldsymbol{A} \cdot \boldsymbol{B})\boldsymbol{C}$ in vector analysis.

Let

$$\boldsymbol{r} = x\boldsymbol{i} + y\boldsymbol{j} + z\boldsymbol{k} \quad \text{and} \quad \boldsymbol{\omega} = \omega_x \boldsymbol{i} + \omega_y \boldsymbol{j} + \omega_z \boldsymbol{k} \,. \tag{2.138}$$

Substituting (2.138) into (2.136), we have

$$\begin{aligned} \boldsymbol{H}_0 = {} & (I_{xx}\omega_x - I_{xy}\omega_y - I_{xz}\omega_z)\,\boldsymbol{i} \\ & + (-I_{xy}\omega_x + I_{yy}\omega_y - I_{yz}\omega_z)\,\boldsymbol{j} \\ & + (-I_{xz}\omega_x - I_{yz}\omega_y + I_{zz}\omega_z)\,\boldsymbol{k} \,, \end{aligned} \tag{2.139}$$

where

$$I_{xx} = \int_m (y^2 + z^2)\, \mathrm{d}m \,, \quad I_{yy} = \int_m (x^2 + z^2)\, \mathrm{d}m \,,$$
$$I_{zz} = \int_m (x^2 + y^2)\, \mathrm{d}m \tag{2.140}$$

are mass moments of inertia about the body axes xyz, and

$$I_{xy} = I_{yx} = \int_m xy\, \mathrm{d}m \,, \; I_{xz} = I_{zx} = \int_m xz\, \mathrm{d}m \,,$$
$$I_{yz} = I_{zy} = \int_m yz\, \mathrm{d}m \tag{2.141}$$

are mass product of inertia about the same axes.

Note that the moment of inertia can be represented as the moment of inertia tensor shown below

$$\boldsymbol{I} = \begin{pmatrix} I_{xx} & -I_{xy} & -I_{xz} \\ -I_{yx} & I_{yy} & -I_{yz} \\ -I_{zx} & -I_{zy} & I_{zz} \end{pmatrix} \,. \tag{2.142}$$

For a pure rotation about a fixed point 0, the moment equation of motion is

$$\boldsymbol{M}_0 = \dot{\boldsymbol{H}}_0 \,. \tag{2.143}$$

If we choose the axes x, y, and z be the principal axes, this results in $I_{xy} = I_{yz} = I_{zx} = 0$. Then the moment equation of motion becomes

$$\begin{aligned} \boldsymbol{M}_0 &= M_x \boldsymbol{i} + M_y \boldsymbol{j} + M_z \boldsymbol{k} \,, \\ M_x &= I_{xx}\dot{\omega}_x + (I_{zz} - I_{yy})\omega_y \omega_z \,, \\ M_y &= I_{yy}\dot{\omega}_y + (I_{xx} - I_{zz})\omega_x \omega_z \,, \\ M_z &= I_{zz}\dot{\omega}_z + (I_{yy} - I_{xx})\omega_x \omega_y \,, \end{aligned} \tag{2.144}$$

which are called Euler's moment equations.

2.2.14 Planar Motion of a Rigid Body

Assuming that the motion takes place in the xy plane, we have $v_z = a_z = 0$, $\omega_x = \dot{\omega}_x = \omega_y = \dot{\omega}_y = 0$, $\omega_z = \omega$, and $I_{xz} = I_{yz} = 0$ due to the small dimension of the body in the z-direction.

For pure translation, i. e., $\omega = \dot{\omega} = 0$, we have

$$F_x = m a_{\mathrm{C}x} \quad \text{and} \quad F_y = m a_{\mathrm{C}y} \,. \tag{2.145}$$

The only moment equation is about the z-axis. The moment equation about the mass center C is $M_{\mathrm{C}z} = 0$.

The kinetic energy is

$$T = \frac{1}{2} m \boldsymbol{v}_{\mathrm{C}} \cdot \boldsymbol{v}_{\mathrm{C}} = \frac{1}{2} m \left(v_{\mathrm{C}x}^2 + v_{\mathrm{C}y}^2 \right) \,, \tag{2.146}$$

where $v_{\mathrm{C}x}$ and $v_{\mathrm{C}y}$ are the Cartesian components of the velocity vector of the mass center.

For pure rotation about a fixed point 0, we have the scalar angular momentum and scalar equation of motion as follows

$$H_0 = I_{zz}\omega \quad \text{and} \quad M_0 = I_{zz}\dot{\omega} \,, \tag{2.147}$$

and the kinetic energy is

$$T = \frac{1}{2} I_{zz} \omega^2 \,. \tag{2.148}$$

Example 2.10: A horizontal bar with a total mass m is hinged at point 0, as depicted in Fig. 2.41. The bar is released from rest. Determine the angular acceleration immediately after release, the reaction force at point 0 at the same time, and the angular velocity of the bar when it passes through the vertical position.

Let us consider the counterclockwise moments and angular motion as positive. From (2.114) and (2.147), we have

$$M_0 = I_{zz}\dot{\omega} = -\frac{1}{6} L m g \,. \tag{2.149a}$$

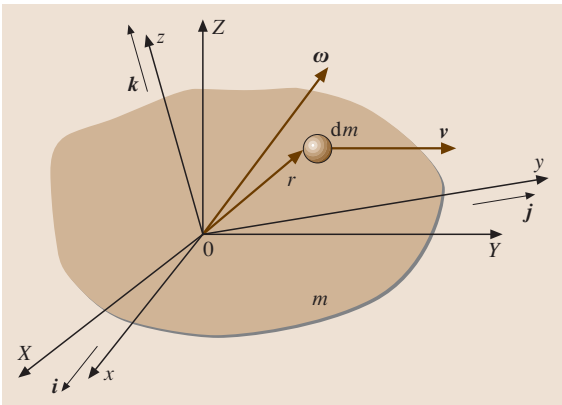


Fig. 2.40 A rotating rigid body

The mass moment of inertia of the bar about point 0 is obtained as

$$I_{zz} = \int_m x^2 \,\mathrm{d}m = \frac{m}{L} \int_{-L/3}^{2L/3} x^2 \,\mathrm{d}x = \frac{1}{9} L^2 m \,. \quad (2.149b)$$

Hence, the angular acceleration immediately after release is

$$\alpha = \dot{\omega} = -\frac{3g}{2L} \,. \quad (2.149c)$$

From (2.53b) and (2.145), we have the reaction force in the y-axis direction only as follows

$$F_y = R - mg = ma_{Cy} = m\frac{1}{6} L\alpha = \frac{-1}{4} mg \,. \quad (2.149d)$$

The reaction force R at point 0 is then obtained as

$$R = \frac{3}{4} mg \,. \quad (2.149e)$$

Both the potential and kinetic energy of the bar at horizontal position are zero. At vertical position, the potential energy becomes $V = -1/6Lmg$. The kinetic energy becomes $T = 1/2 I_{zz}\omega^2 = 1/18 mL^2\omega^2$. By applying the law of conservation of energy, we have

$$T + V = 0 \quad \text{and} \quad \omega = -\sqrt{\frac{3g}{L}} \,, \quad (2.149f)$$

where te negative sign indicates an angular velocity in the clockwise direction.

2.2.15 General Case of Planar Motion

By using the mass center C, the moment equation has the scalar form $M_c = I_c\alpha$, where I_c is the mass moment

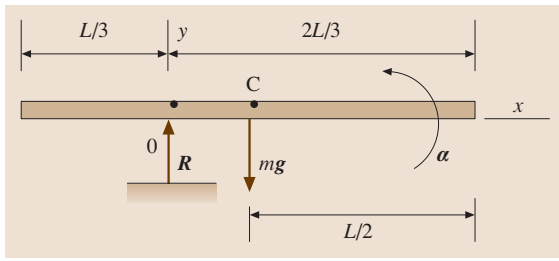


Fig. 2.41 A uniform bar

of inertia about an axis normal to the plane of motion that passing through C. The kinetic energy consists of the translation of C and rotation about C as follows

$$T = \frac{1}{2} m\left(v_{Cx}^2 + v_{Cy}^2\right) + \frac{1}{2} I_C \omega^2 \,. \quad (2.150)$$

Let us consider the system shown in Fig. 2.42 and write a moment about the arbitrary point A as

$$\boldsymbol{M}_A = \int \boldsymbol{\rho}_A \times \mathrm{d}\boldsymbol{F} = \int (\boldsymbol{\rho}_{AC} + \boldsymbol{\rho}) \times \mathrm{d}\boldsymbol{F} = \boldsymbol{\rho}_{AC} \times \boldsymbol{F} + \boldsymbol{M}_C \,, \quad (2.151)$$

or

$$\boldsymbol{M}_A = \boldsymbol{\rho}_{AC} \times m\boldsymbol{a}_C + \boldsymbol{M}_C \quad \text{and} \quad \boldsymbol{M}_C = I_C\boldsymbol{\alpha} \,. \quad (2.152)$$

The acceleration at the mass center C can be written as

$$\boldsymbol{a}_C = \boldsymbol{a}_A + \boldsymbol{a}_{C/A} \,. \quad (2.153)$$

From (2.82), the acceleration reduces to

$$\boldsymbol{a}_C = \boldsymbol{a}_A - \omega^2 \boldsymbol{\rho}_{AC} + \boldsymbol{\alpha} \times \boldsymbol{\rho}_{AC} \,. \quad (2.154)$$

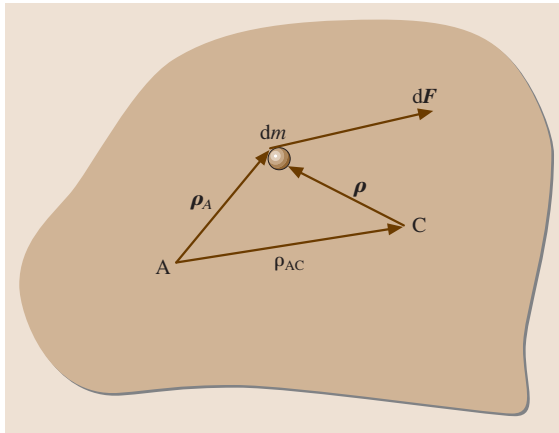


Fig. 2.42 Rigid-body planar motion

Substituting (2.154) into (2.152) we have

$$\boldsymbol{M}_A = \boldsymbol{\rho}_{AC} \times m\boldsymbol{a}_C + I_A \boldsymbol{\alpha} \,. \tag{2.155}$$

where $I_A = (I_C + m\rho_{AC}^2)$ is the mass moment of inertia of the body about point A.

Example 2.11: A disk of radius R is originally at rest at point 1. It rolls without slipping down to point 2 as depicted in Fig. 2.43. Calculate the velocity at point 2.

The kinetic energy at point 2 is

$$T_2 = \frac{1}{2} m v_2^2 + \frac{1}{2} I_C \omega_2^2 \,. \tag{2.156a}$$

Since $v_2 = -R\omega_2$ and $I_C = mR^2/2$, we have

$$T_2 = \frac{3}{4} m v_2^2 \,. \tag{2.156b}$$

Conservation of energy then yields

$$V_2 + T_2 = -mgh + \frac{3}{4} m v_2^2 = 0 \quad \text{and}$$
$$v_2 = \sqrt{4gh/3} \,. \tag{2.156c}$$

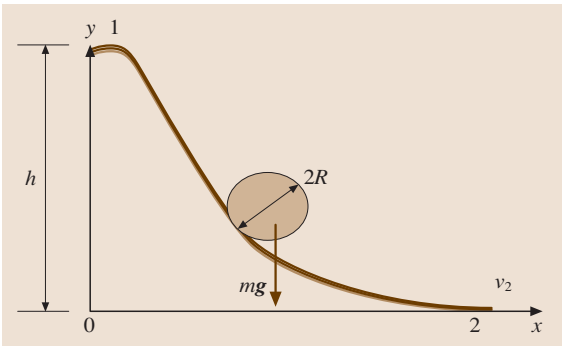


Fig. 2.43 A rolling ball

2.2.16 Rotation About a Fixed Axis

Let us consider the rigid body of Fig. 2.40 and assume that the only motion takes place about the fixed z-axis, i. e., $\omega_x = \dot{\omega}_x = \omega_y = \dot{\omega}_y = 0$, $\omega_z = \omega$. The fixed origin 0 is also on the fixed axis. The moment equations become

$$\begin{aligned} M_x &= -I_{xz}\dot{\omega} + I_{yz}\omega^2 \,, \\ M_y &= -I_{yz}\dot{\omega} - I_{xz}\omega^2 \,, \\ M_z &= I_{zz}\dot{\omega} \,. \end{aligned} \tag{2.157}$$

Example 2.12: A thin disk with radius R and mass m is shown in Fig. 2.44. The normal to the disk makes an angle β with respect to the shaft. The disk rotates with $\omega_z = \omega = \text{const}$. Determine the bearing forces at points A and B.

From Fig. 2.44a, axes XYZ are inertial and axes xyz are body axes, where z is along the shaft and x is embedded in the disk. Fig. 2.44b depicts the body axes, xyz and the principle axes, $x'y'z'$. The relationship between these two set of axes are

$$\begin{aligned} &x = x' \,, \quad y = y' \cos\beta + z' \sin\beta \,, \\ &z = -y' \sin\beta + z' \cos\beta \,. \end{aligned} \tag{2.158a}$$

The inertia products I_{xz} and I_{yz} of the axes xyz are

$$\begin{aligned} I_{xz} &= \int_m xz \, \mathrm{d}m = \int_m x'(-y' \sin\beta + z' \cos\beta) \, \mathrm{d}m \\ &= -\sin\beta I_{x'y'} + \cos\beta I_{x'z'} = 0 \end{aligned} \tag{2.158b}$$

and

$$\begin{aligned} I_{yz} &= \int_m yz \, \mathrm{d}m \\ &= \int_m (y' \cos\beta + z' \sin\beta) \\ &\quad \times (-y' \sin\beta + z' \cos\beta) \, \mathrm{d}m \\ &= \sin\beta \cos\beta (I_{y'y'} - I_{z'z'}) \\ &\quad + (\cos^2\beta - \sin^2\beta) I_{y'z'} \,. \end{aligned} \tag{2.158c}$$

Note that because $x'y'z'$ are the principle axes, the products of inertia are zero. The moments of inertia of the disk are

$$I_{x'x'} = I_{y'y'} = \frac{1}{4} m R^2 \,, \quad I_{z'z'} = \frac{1}{2} m R^2 \,. \tag{2.158d}$$

Hence,

$$I_{yz} = -\frac{1}{4} m R^2 \sin\beta \cos\beta \,. \tag{2.158e}$$

Substituting (2.158b) and (2.158e) into (2.157), we obtain

$$M_x = -\frac{1}{4} m R^2 \omega^2 \sin\beta \cos\beta \,, \quad M_y = 0, \quad M_z = 0 \,. \tag{2.158f}$$

The moment components along the X- and Y-axes are

$$M_X = M_x \cos\gamma \quad \text{and} \quad M_Y = M_x \sin\gamma \,. \tag{2.158g}$$

Since the acceleration of the mass center is zero, the force equations along the X- and Y-axes are

$$\begin{aligned} &F_X = R_{AX} + R_{BX} = 0 \quad \text{and} \\ &F_Y = R_{AY} + R_{BY} - mg = 0 \,. \end{aligned} \tag{2.158h}$$

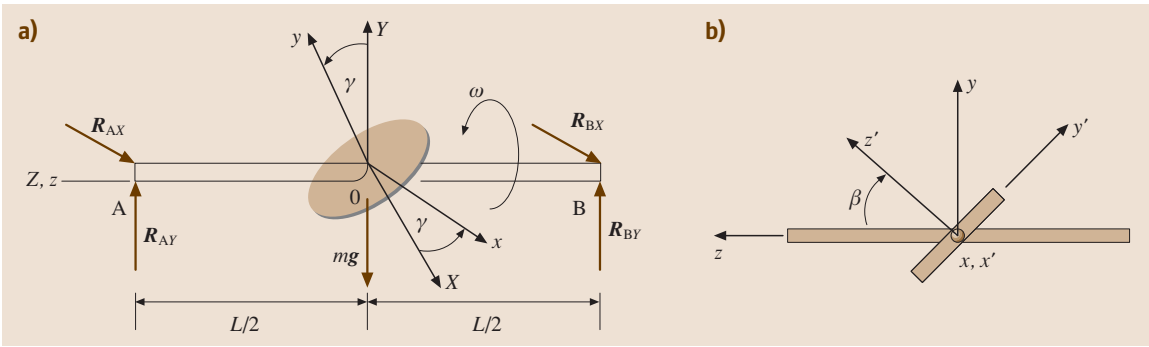


Fig. 2.44a,b A rotating disk: (a) the axes XYZ are inertial, (b) the axes xyz are body axes

From the reaction forces, the moment equations about 0 yield

$$M_X = \frac{L}{2}(R_{BY} - R_{AY}) = -\frac{1}{4}mR^2\omega^2 \sin\beta\cos\beta\cos\gamma\,,$$
$$M_Y = \frac{L}{2}(R_{AX} - R_{BX}) = -\frac{1}{4}mR^2\omega^2 \sin\beta\cos\beta\sin\gamma\,. \qquad (2.158i)$$

Because $\gamma = \omega t$, (2.158h) and (2.158i) yield

$$R_{BX} = -R_{AX} = \frac{1}{4L}mR^2\omega^2 \sin\beta\cos\beta\sin\omega t\,,$$
$$R_{AY} = \frac{mg}{2} + \frac{1}{4L}mR^2\omega^2 \sin\beta\cos\beta\cos\omega t\,,$$
$$R_{BY} = \frac{mg}{2} - \frac{1}{4L}mR^2\omega^2 \sin\beta\cos\beta\cos\omega t\,. \qquad (2.158j)$$

Therefore, in addition to the static bearing forces equal to half the weight, there are dynamic bearing forces that vary harmonically with a frequency equal to the spin frequency ω. These dynamic bearing forces will wear out the bearing.

2.2.17 Lagrange's Equations of Motion for Linear Systems

Lagrange's equations can be applied to the derivation of the equations of motion for a linear n-degree-of-freedom dynamic system. By extending the concept of Sect. 2.2.12, (2.134) can be rewritten as follows

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial T}{\partial \dot{q}_i}\right) - \frac{\partial T}{\partial q_i} + \frac{\partial D}{\partial \dot{q}_i} + \frac{\partial V}{\partial q_i} = Q_i\,, \quad i = 1, 2, \ldots, n\,, \qquad (2.159)$$

where D, the dissipation function due to the damping force of viscous type, is defined as

$$D = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}\dot{q}_i\dot{q}_j \quad \text{and}$$
$$\frac{\partial D}{\partial \dot{q}_i} = \sum_{j=1}^{n} c_{ij}\dot{q}_j\,, \qquad (2.160)$$

where the c_{ij} are known as the damping coefficients.

Equation (2.159) can be rewritten in a compact matrix form as

$$M\ddot{\boldsymbol{q}}(t) + C\dot{\boldsymbol{q}}(t) + K\boldsymbol{q}(t) = \boldsymbol{Q}(t)\,, \qquad (2.161)$$

where M is the mass matrix, C is the damping matrix, and K is the stiffness matrix. All three are $n \times n$ symmetric matrices.

$$\boldsymbol{q}(t) = [q_1(t)q_2(t), \ldots, q_n(t)]^{\mathrm{T}} \quad \text{and}$$
$$\boldsymbol{Q}(t) = [Q_1(t), Q_2(t), \ldots, Q_n(t)]^{\mathrm{T}} \qquad (2.162)$$

Equation (2.162) defines the n-dimensional generalized displacement vector and generalized force vector.

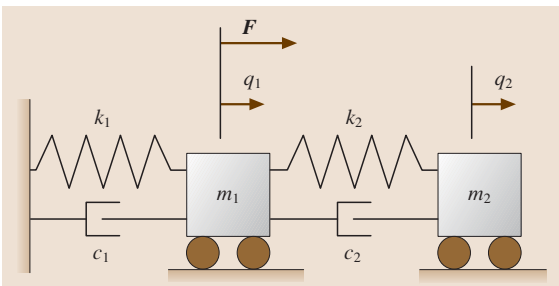


Fig. 2.45 Spring–mass–damping system

The kinetic energy T, dissipation function D, and the potential energy V can be expressed as

$$T = \frac{1}{2}\dot{\boldsymbol{q}}^{\mathrm{T}}(t)M\dot{\boldsymbol{q}}(t)$$
$$D = \frac{1}{2}\dot{\boldsymbol{q}}^{\mathrm{T}}(t)C\dot{\boldsymbol{q}}(t)$$
$$V = \frac{1}{2}\boldsymbol{q}^{\mathrm{T}}(t)K\boldsymbol{q}(t)\,. \qquad (2.163)$$

Example 2.13: Derive the equation of motion of the system shown in Fig. 2.45. Use q_1 and q_2 as independent coordinates.

The kinetic energy T, dissipation function D, and the potential energy V are

$$T = \frac{1}{2}\left(m_1\dot{q}_1^2 + m_2\dot{q}_2^2\right)$$
$$= \frac{1}{2}\begin{pmatrix}\dot{q}_1\\ \dot{q}_2\end{pmatrix}^{\mathrm{T}}\begin{pmatrix}m_1 & 0\\ 0 & m_2\end{pmatrix}\begin{pmatrix}\dot{q}_1\\ \dot{q}_2\end{pmatrix},$$
$$D = \frac{1}{2}\left[c_1\dot{q}_1^2 + c_2(\dot{q}_2 - \dot{q}_1)^2\right]$$
$$= \frac{1}{2}\begin{pmatrix}\dot{q}_1\\ \dot{q}_2\end{pmatrix}^{\mathrm{T}}\begin{pmatrix}c_1 + c_2 & -c_2\\ -c_2 & c_2\end{pmatrix}\begin{pmatrix}\dot{q}_1\\ \dot{q}_2\end{pmatrix},$$
$$V = \frac{1}{2}\left[k_1 q_1^2 + k_2(q_2 - q_1)^2\right]$$
$$= \frac{1}{2}\begin{pmatrix}q_1\\ q_2\end{pmatrix}^{\mathrm{T}}\begin{pmatrix}k_1 + k_2 & -k_2\\ -k_2 & k_2\end{pmatrix}\begin{pmatrix}q_1\\ q_2\end{pmatrix}. \qquad (2.164a)$$

The equation of motion of the two-degree-of-freedom dynamic system is obtained as follows

$$\begin{pmatrix}m_1 & 0\\ 0 & m_2\end{pmatrix}\begin{pmatrix}\ddot{q}_1\\ \ddot{q}_2\end{pmatrix} + \begin{pmatrix}c_1 + c_2 & -c_2\\ -c_2 & c_2\end{pmatrix}\begin{pmatrix}\dot{q}_1\\ \dot{q}_2\end{pmatrix} + \begin{pmatrix}k_1 + k_2 & -k_2\\ -k_2 & k_2\end{pmatrix}\begin{pmatrix}q_1\\ q_2\end{pmatrix} = \begin{pmatrix}F\\ 0\end{pmatrix}. \qquad (2.164b)$$

This second-order differential equation can be solved and time- and frequency-domain responses can be obtained. Furthermore, this type of problem can be treated as a spring–mass-damping vibration system. Additionally, modern control theory can be introduced to this type of dynamic systems with feedback loops to obtain the desired time- or frequency-domain response.

References

2.1 Y.C. Fung: *A First Course in Continuum Mechanics* (Prentice-Hall, Old Tappan 1969) p. 2

2.2 J.L. Meriam, L.G. Kraige: Engineering Mechanics. In: *Statics*, Vol. 1 (Wiley, New York 2002) p. 4

2.3 F.P. Beer, E.R. Johnston Jr., E.R. Eisenberg: *Vector Mechanics for Engineers – Statics* (McgGraw Hill, New York 2004) pp. 36, 159

2.4 W.F. Rilet, L.D. Sturges: *Engineering Mechanics – Statics* (Wiley, New Jersey 1993) p. 263

2.5 F.P. Beer, E.R. Johnson Jr., J.T. DeWolf: *Vector Mechanics for Engineers: Statics* (McGraw-Hill, New York 2006)

2.6 R.C. Hibbeler: *Engineering Mechanics – Dynamics*, 11th edn. (Prentice-Hall, New Jersey 2006)

2.7 J.L. Meriam, L.G. Kraige: *Engineering Mechanics: Dynamics*, 6th edn. (Wiley, New York 2006)

2.8 F.P. Beer: *Vector Mechanics for Engineers: Dynamics* (McGraw-Hill, New York 2005)

2.9 A. Bedford, W. Fowler: *Engineering Mechanics – Statistics* (Prentice-Hall, Old Tappan 2005) p. 448

Part B Applications

Part B Applications in Mechanical Engineering

3 Materials Science and Engineering

Jens Freudenberger, Dresden, Germany
Joachim Göllner, Magdeburg, Germany
Martin Heilmaier, Darmstadt, Germany
Gerhard Mook, Magdeburg, Germany
Holger Saage, Landshut, Germany
Vivek Srivastava, Navi Mumbai, India
Ulrich Wendt, Magdeburg, Germany

4 Thermodynamics

Frank Dammel, Darmstadt, Germany
Jay M. Ochterbeck, Clemson, USA
Peter Stephan, Darmstadt, Germany

5 Tribology

Ludger Deters, Magdeburg, Germany

6 Design of Machine Elements

Oleg P. Lelikov, Moscow, Russia

7 Manufacturing Engineering

Thomas Böllinghaus, Berlin, Germany
Gerry Byrne, Belfield, Dublin 4, Ireland
Boris Ilich Cherpakov (deceased)
Edward Chlebus, Wrocław, Poland
Carl E. Cross, Berlin, Germany
Berend Denkena, Garbsen, Germany
Ulrich Dilthey, Aachen, Germany
Takeshi Hatsuzawa, Yokohama, Japan
Klaus Herfurth, Langenfeld, Germany
Horst Herold (deceased)
Andrew Kaldos, Bebington, UK
Thomas Kannengiesser, Berlin, Germany
Michail Karpenko, Manukau City, New Zealand
Bernhard Karpuschewski, Magdeburg, Germany
Manuel Marya, Rosharon, USA
Surendar K. Marya, Nantes, France
Klaus-Jürgen Matthes, Chemnitz, Germany
Klaus Middeldorf, Düsseldorf, Germany
Joao Fernando G. Oliveira, São Carlos, Brazil
Jörg Pieschel, Magdeburg, Germany
Didier M. Priem, Nantes, France
Frank Riedel, Chemnitz, Germany
Markus Schleser, Aachen, Germany
A. Erman Tekkaya, Ankara, Turkey
Marcel Todtermuschke, Chemnitz, Germany
Anatole Vereschaka, Moscow, Russia
Detlef von Hofe, Krefeld, Germany
Nikolaus Wagner, Aachen, Germany
Johannes Wodara, Magdeburg, Germany
Klaus Woeste, Aachen, Germany

8 Measuring and Quality Control

Norge I. Coello Machado, Santa Clara, Cuba
Shuichi Sakamoto, Niigata, Japan
Steffen Wengler, Magdeburg, Germany
Lutz Wisweh, Magdeburg, Germany

9 Engineering Design

Alois Breiing, Zurich, Switzerland
Frank Engelmann, Jena, Germany
Timothy Gutowski, Cambridge, USA

contd.

10 Piston Machines

Vince Piacenti, Farmington Hills, USA
Helmut Tschoeke, Magdeburg, Germany
Jon H. Van Gerpen, Moscow, USA

11 Pressure Vessels and Heat Exchangers

Ajay Mathur, New Delhi, India

12 Turbomachinery

Meinhard T. Schobeiri, College Station, USA

13 Transport Systems

Gritt Ahrens, Sindelfingen, Germany
Torsten Dellmann, Aachen, Germany
Stefan Gies, Aachen, Germany
Markus Hecht, Berlin, Germany
Hamid Hefazi, Long Beach, USA
Rolf Henke, Aachen, Germany
Stefan Pischinger, Aachen, Germany
Roger Schaufele, Long Beach, USA
Oliver Tegel, Weissach, Germany

14 Construction Machinery

Eugeniusz Budny, Warsaw, Poland
Mirosław Chłosta, Warsaw, Poland
Henning Jürgen Meyer, Berlin, Germany
Mirosław J. Skibniewski, College Park, USA

15 Enterprise Organization and Operation

Francesco Costanzo, Pomigliano (NA), Italy
Yuichi Kanda, Kawagoe-City, Japan
Toshiaki Kimura, Tokyo, Japan
Hermann Kühnle, Magdeburg, Germany
Bruno Lisanti, Lonate Pozzolo (VA), Italy
Jagjit Singh Srai, Cambridge, UK
Klaus-Dieter Thoben, Bremen, Germany
Bernd Wilhelm, Wolfsburg, Germany
Patrick M. Williams, Bristol, UK

3. Materials Science and Engineering

Jens Freudenberger, Joachim Göllner, Martin Heilmaier, Gerhard Mook, Holger Saage, Vivek Srivastava, Ulrich Wendt

The chapter is structured into the following main parts. After a short introduction which addresses the term *materials* as it is used in mechanical engineering and sorts out other matters for the sake of space, the first main section, Sect. 3.1, describes the fundamentals of *atomic structure and microstructure* of materials (as defined in the introduction). The following Sects. 3.3, 3.4, 3.5, 3.6 deal with the most important *properties and testing methods* of materials from the viewpoint of mechanical engineers. The last and largest Sect. 3.7 is devoted to the most commonly used *materials in mechanical engineering*.

Materials science and technology is still a relatively young scientific discipline with its roots dating back about half a century. Emerging from the schools in metal physics in Cambridge, Göttingen, Oxford, and Stuttgart amongst others, it is, in essence, a truly interdisciplinary field where people with a classical education in natural sciences, e.g., in (solid-state) physics and (physical) chemistry to mention a few, but also with a solid engineering background, e.g., in chemical or mechanical engineering, come together to develop new materials with improved properties for the ever-increasing demand of our society. Bearing this in mind, the present chapter is not intended to address all these aspects and recent developments in depth; thus, it is

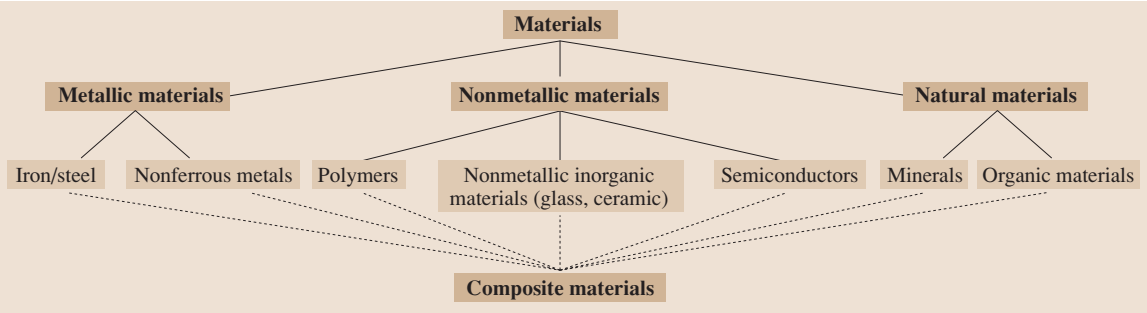


Fig. 3.1 Scheme for the classification of materials

far from being complete. Rather, the authors wanted to give both the novice and the expert a broad but up-to-date overview of those topics of *materials science and technology* that are relevant to the field of *mechanical engineering*.

Though liquids and gases are frequently used in mechanical engineering, e.g., lubricants to reduce friction in tribological applications (Chap. 5) or oils in combustion engines (Chap. 10) the most commonly used definition of materials in mechanical engineering focuses on those in the solid state during their technical application. Still, this broad spectrum requires a further classification, as illustrated in Fig. 3.1, which takes into account the traditional special role of *metallic materials* in mechanical engineering. The term *metallic materials* is used here not only for pure metals, i. e., the metallic elements of the periodic table, but also for *alloys*, which can be created by solving other chemical elements within the crystallographic structure of the base (matrix) element (Sect. 3.2). Metals contrast with *nonmetallic materials* (the second group in Fig. 3.1, in essence comprising *polymers*, *ceramics* and *glasses*, and semiconductors) because of their particular physical properties (see Sect. 3.2.2 for an overview and [3.1] for details). Most importantly, their electrical conductivity is usually several orders of magnitudes larger than in nonmetallic materials while characteristically decreasing with increasing temperature. Further typical metallic properties are their high thermal conductivity, nontransparency to visible light (in bulk form), and a shiny surface, all of which stem from the special electronic structure, i. e., from the *metallic bonding*. For details regarding the peculiarities of the different types of bonding (i. e., *ionic*, *covalent*, and *van der Waals* types) the interested reader should refer to textbooks on solid-state physics or physical chemistry [3.2, 3]. The scheme in Fig. 3.1 is completed with *natural materials*, which can be further subdivided into *minerals* and *organic (natural) materials*. Representatives of the former group are, e.g., *stone* (including *precious stones* such as *diamond*) and *asbestos*, while the latter is essentially comprises *wood* and *rubber*. However, due to their limited importance for mechanical engineering in general, natural materials will not be addressed in the remainder of this chapter.

The group of metallic materials is usually divided into *iron and steel* and *nonferrous metals*. Iron and steel – in essence an alloy of Fe and C – are still the most important structural materials by far to deal with in mechanical engineering. The reason for this importance lies mainly in the availability of raw materials, the sophisticated processing, and the possibility of tuning the mechanical properties within an extremely wide spectrum. Steels will be highlighted in Sect. 3.3.1. However, nonferrous metals are gaining increasing importance in mechanical engineering mainly due to specific properties stemming from their chemistry. Some of the more important ones which will be discussed in more detail later are *light metals* and alloys of aluminum (Sect. 3.3.2), magnesium (Sect. 3.3), titanium (Sect. 3.4), *nickel-based alloys* for high-temperature applications (Sect. 3.5), and *copper* and its alloys for conducting applications (Sect. 3.6). For precious metals (Ag, Au, Pt, Pd) and refractory metals (Mo, W, Nb, Ta) and their properties and fields of applications the interested reader is referred to [3.1].

Of the nonmetallic materials polymers is becoming increasingly important in mechanical engineering, mainly because of its low specific weight and ease of manufacturing. They will be treated in Sect. 3.7.7. Nonmetallic anorganic materials, i. e., glasses and ceramics, are described in Sect. 3.7.9. Finally, combining two or more of the subclasses described before and in Fig. 3.1 leads to the emerging field of *composite materials*. These will be highlighted in a few examples in Sect. 3.3.

3.1 Atomic Structure and Microstructure

3.1.1 Atomic Order in Solid State

While many physical and/or chemical problems should be considered on the atomic level, it may be satisfactory for mechanical engineers to remain on the microstructural level. Therefore, we will discuss in this section the different possibilities of (ideal) atomic arrangements, whereas we introduce the so-called *lattice defects* in the subsequent section. Then, three main categories of atomic arrangement can be distinguished (Fig. 3.2):

- No order (or disordered state), which is the case for inert gases such as argon (Fig. 3.2a), where the interaction between the single atoms is essentially limited to random collisions. Since we focus on materials in the solid state, this will be disregarded in the following.
- Short-range order (SRO) over only a few atomic distances, which can be observed for polar molecules such as water (Fig. 3.2b), but also in polymers (e.g., polyethylene) and glasses; e.g., silica is built of chains of tetrahedrons with a (central) Si atom surrounded by four oxygen atoms (Fig. 3.2c). These materials are, thus, called *amorphous solids* or, in view of the similarity to polar liquids, *undercooled liquids*.
- Long-range order (LRO) requires the atoms to be arranged on a periodic *crystal lattice* (*crystallos* is Greek for *ice*) with – in principle – infinite extension in three-dimensional space where the atoms sit on certain lattice points in such a way that the next neighbor situation is the same for every atom under consideration. An example of a *primitive cubic lattice* is shown in Fig. 3.2d. Such an ideal atomic arrangement is also called *single crystal*. Then, the smallest possible three-dimensional geometrical unit able to reproduce the lattice structure is called the *unit cell* of the corresponding crystal structure with the three unit vectors being the *lattice constants* of the crystal structure. This equilibrium distance of two atoms can then be considered as the result of a superposition of an attractive and a repulsive potential between these atoms.

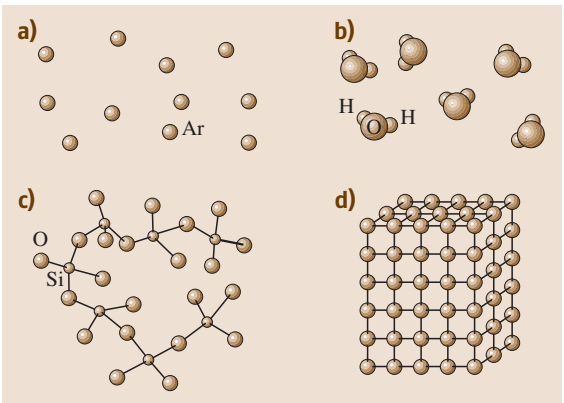


Fig. 3.2a–d Categories of atomic arrangement in materials: **(a)** inert gases have no order, **(b)** and **(c)** polar liquids and amorphous solids show short-range order, **(d)** crystalline materials possess long-range order (of infinite extension)

Amorphous Structures

As pointed out in the previous subsection two important classes of materials, namely glasses and many polymers, exist in the solid state without possessing a long-periodic crystallographic lattice structure: they are amorphous solids, thus, having SRO. Figure 3.3 elucidates how these structures may form upon cooling

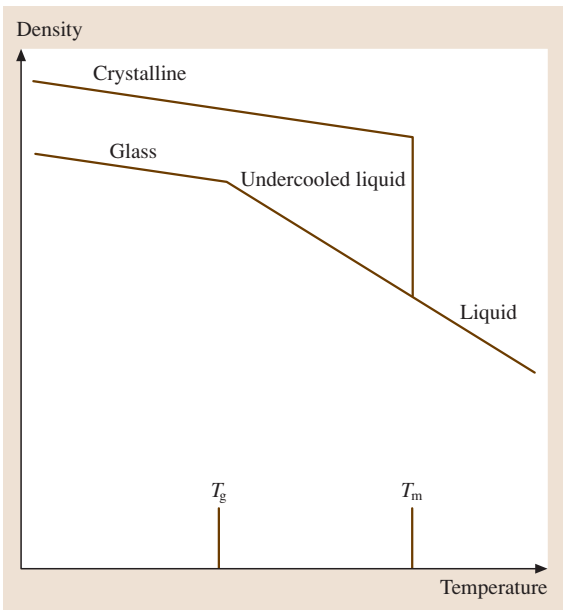


Fig. 3.3 Cooling of a silica melt: crystallization at T_m leads to an abrupt increase in density; if crystallization is suppressed, however, the undercooled melt transforms into a glassy state at T_g and the density–temperature curve shows a bend

from the melt: usually, a liquid starts to form a crystalline solid when it is cooled below the melting point T_m of the material due to crystal nucleation and growth (see Sect. 3.2.3 for details). The crystallization event is accompanied with an abrupt jump in density. If, however, the kinetics for this transition into solid state is too sluggish or the cooling rate is too high, crystallization may be completely suppressed and an undercooled liquid is formed, which eventually transforms below the glass-transition temperature T_g into a glassy state. Note, that the transition from the undercooled liquid into a glassy state (with both structures being *amorphous* and possessing SRO) is a second-order thermodynamic reaction. This expresses itself in Fig. 3.3 as a bend of the curve such that both the crystalline and the glassy state show an identical temperature dependence of density.

An important conclusion can be drawn from Fig. 3.3. Since the glassy state exhibits the lower density, it is in a thermodynamically metastable state (of higher free energy G, Sect. 3.2.3) and will eventually transform into the crystalline structure, i. e., into the thermodynamically stable state (of minimal free energy G_{min}).

Like glasses and polymers some metals can also be solidified into an amorphous structure. However, while simple binary and ternary systems can be produced only in the form of thin metallic ribbons via rapid solidification (requiring cooling rates of up to 10^6 K/s; see [3.4] for a review), multicomponent metallic glassy alloys in bulk form produced by slow cooling from the melt have attracted widespread interest ranging from scientific curiosity about their structure and resulting properties to technological aspects of their preparation and potential applications. Readers interested in the outstanding properties of these new emerging class of advanced metallic materials may be referred to relevant literature [3.5–8].

Crystal Structures

As early as the 19th century the French scientist *Bravais* demonstrated that in three-dimensional space the seven major crystallographic unit cells may be better classified into 14 *Bravais lattices* (Fig. 3.4). The resulting geometric relations between the crystal axes are tabulated in Table 3.1. However, of those listed, most metallic materials used in mechanical engineering crystallize in hexagonal or in cubic form, i. e., in crystal lattice structures with high symmetry.

The most important characteristics of these crystal structures are summarized in Table 3.2. In essence they

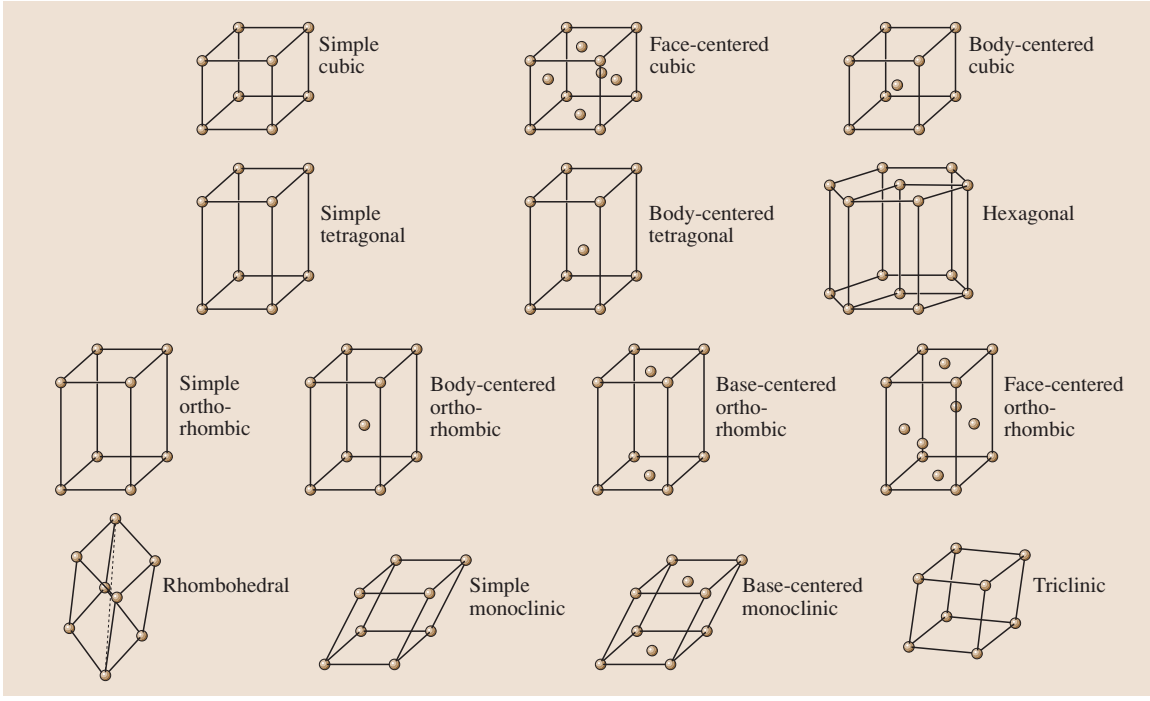


Fig. 3.4 The seven crystal systems and the 14 types of unit cells (Bravais lattices)

Table 3.1 Characteristic relations between crystallographic axes (lattice constants) and angles within the seven crystal systems

Structure	Lattice constants	Angles between axes
Cubic	$a = b = c$	All angles equal 90°
Tetragonal	$a = b \neq c$	All angles equal 90°
Orthorombic	$a \neq b \neq c$	All angles equal 90°
Hexagonal	$a = b \neq c$	Two angles equal 90°. One angle equals 120°
Rhombohedric	$a = b = c$	All angles are equal and none equals 90°
Monoclinic	$a \neq b \neq c$	Two angles equal 90°. One angle (β) is not equal to 90°
Triclinic	$a \neq b \neq c$	All angles are different and none equals 90°

can be distinguished by the number of atoms per unit cell (NA), the coordination number (i. e., the number of nearest neighbors, CN) and the packing density (PD). The latter is defined as $\mathrm{PD} = \mathrm{NA}/V_{\mathrm{uc}}$ (with V_{uc} being the volume of the unit cell) and ranges between 0 and 1.

A comparison of the different crystal structures in Table 3.2 yields that:

1. For the cubic systems, both the coordination number and the packing density increase with increasing number of atoms in the unit cell.
2. CN and PD are identical for face-centered cubic (fcc) and hexagonal crystals, if the ratio of the c- and the a-axis in the hexagonal system is 1.633 (which is nearly the case for the metallic elements Ti, Mg, and Co).

Then, these systems are called *hexagonally closed packed* (hcp) and they can be distinguished from the fcc structures only by the different stacking sequence, which is ABAB... for hcp along the c-axis and ABCABC... for fcc along the space diagonal ⟨111⟩ (in crystallography it is convenient to use the Miller indices for describing lattice vectors and planes; in cubic systems the space diagonal is a ⟨111⟩ direction [3.3]).

Table 3.2 Characteristics of cubic and hexagonal unit cells. NA is the number of atoms per unit cell on regular lattice points, CN is the coordination number, PD is the packing density

Structure	NA	CN	PD
Simple cubic	1	6	0.52
Body-centered cubic (bcc)	2	8	0.68
Face-centered cubic (fcc)	3	12	0.74
Hexagonal	6	12	0.74 (if $c/a = 1.633$)

It is obvious from the comparison in Table 3.2 that even the closed packed crystal structures with space filling of 74% are still far from being fully dense. The open spaces left between regular lattice sites are called *interstitial sites*. They may be filled by atomic species that are significantly smaller than the matrix atoms which build the regular crystal lattice. For cubic systems three different interstitial sites can be distinguished depending on the crystal structure and on the ratio of the radii of the (foreign) interstitial atom r_{ia} and the matrix atom r_{m}, respectively (Fig. 3.5). For relatively large interstitial atoms and $r_{\mathrm{ia}}/r_{\mathrm{m}} = 0.732\ldots1$ these interstitials will likely occupy *cubic interstitial lattice sites* with $\mathrm{CN} = 8$. *Octahedral interstitial sites* with $\mathrm{CN} = 6$ may be filled when $r_{\mathrm{ia}}/r_{\mathrm{m}} = 0.414\ldots0.732$. *Tetrahedral interstitial sites* with $\mathrm{CN} = 4$ provide the least free space and, hence, $r_{\mathrm{ia}}/r_{\mathrm{m}} = 0.225\ldots0.414$. This is technically most relevant since the interstitial carbon atom in the Fe – C alloy system (i. e., *steel*) obeys these empirical rules and favors octahedral sites (largest holes) in the fcc structure, whereas it occupies tetrahedral sites in the body-centered cubic (bcc) structure. The incorporation of carbon atoms at interstitial sites leads to significant elastic lattice distortion and, thus, readily explains the increased strength of steels as compared with pure iron.

Polymorphism

If materials exist in more than one crystal structure depending on temperature and/or pressure they are called *polymorphic*. A more specific term applicable for pure elements is *allotropy*. Two prominent representatives for metallic materials show polymorphism:

1. Iron and steel transforms from the low-T *ferritic* bcc structure into a higher-temperature *austenitic* fcc structure and back to a bcc structure called *δ-ferrite* close to the melting point.
2. Titanium undergoes a transformation from α-hcp structure at low T to a β-bcc structure at high T.

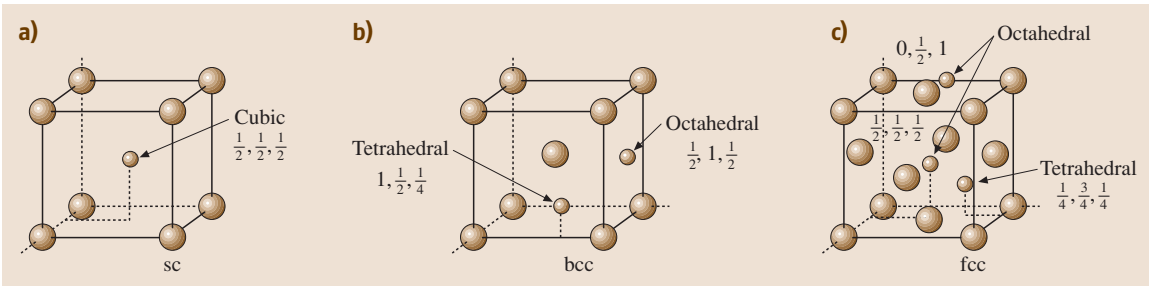


Fig. 3.5a–c Locations of interstitial sites within cubic unit cells: **(a)** simple cubic, **(b)** body-centered cubic (bcc), **(c)** face-centered cubic. The numbers denote the lattice positions of the interstitial sites

Both transformation reactions form the basis for the heat treatment of steels and titanium alloys and, thus, the potential for widely adjusting the properties of these alloy systems; see Sect. 3.3.1 and Sect. 3.7.4, respectively. A further example of polymorphism is the covalently bonded element carbon, which exists in a hexagonal structure of two-dimensional layers called *graphite*, as *diamond* with the open diamond cubic crystal structure (with a PD of only 0.34) and, as recently discovered, in the form of a hollow C_{60} *spherical molecule* called *fullerenes* or bucky balls (the scientists R. F. Curl, H. W. Kroto, and R. E. Smalley received the Noble prize in chemistry in 1996 for this discovery).

Crystal Structures with More Than One Atomic Species

In most cases, technically relevant materials consist of more than one atomic species. Depending on how the different elements mixed with each other, one can classify them into:

1. *Solid solutions*, where the different atomic elements are randomly mixed and placed on the lattice points of the crystal structure. Two subcases may be distinguished:

- Substitutional solid solutions, where the atomic species occupy regular lattice points in a random manner (see also Fig. 3.7c,d). According to *Hume-Rothery* [3.3, 10] this occurs when the elements (e.g., A and B) have comparable atomic radii ($\delta = (r_B - r_A)/r_A < 15\%$) and crystallize in the same structure. A prominent example is the binary alloy system Cu – Ni, which exhibits unlimited mutual solubility.
- Interstitial solid solution, where the smaller atomic species occupy interstitial sites because of the too large a difference in atomic radii between them and the matrix atoms ($\delta > 15\%$) (Fig. 3.7b). The most prominent example is again C in Fe, but this can be generalized for gaseous impurities such as H, O, and N in metals. In contrast to the former case the maximum solubility is much smaller, typically less than 1 at.%, which means that in the case of a bcc lattice only about every 50th unit cell hosts one interstitially solved atom:

2. *Ordered* crystal structures, where the ratio between the atomic species is fixed to small integer numbers. This is called *stoichiometry*. Two categories may be

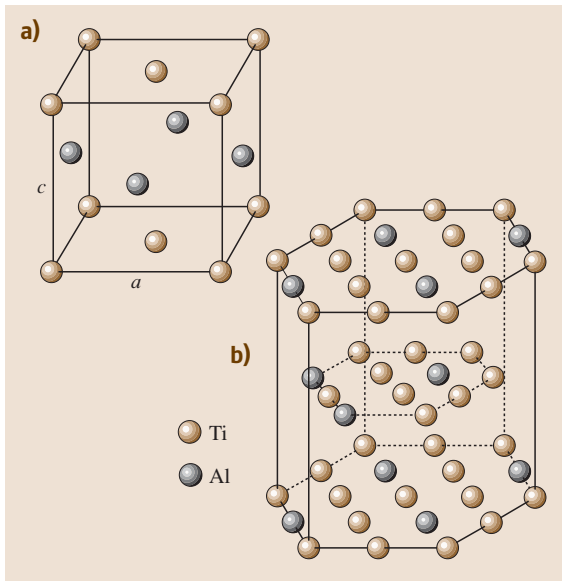


Fig. 3.6a,b Some unit cells of technically relevant intermetallic phases within the binary Ti–Al system: **(a)** γ-TiAl with (tetragonally distorted) $L1_0$ crystal structure, **(b)** α_2-Ti_3Al with hexagonal $D0_{19}$ crystal structure (Pearson symbols [3.9])

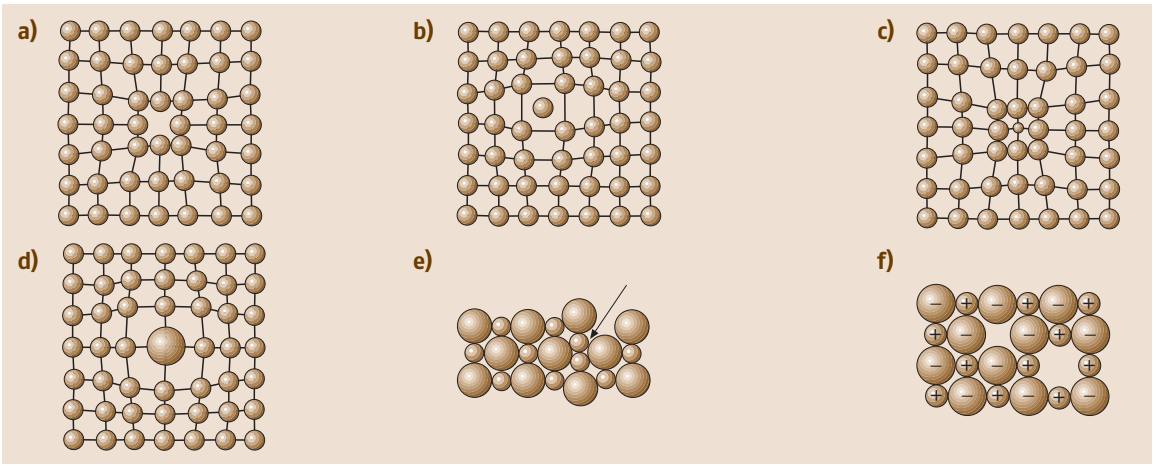


Fig. 3.7a–f Overview of the various types of point defects: **(a)** vacancy, **(b)** interstitial atom, **(c)** and **(d)** smaller and larger substitutional atom, **(e)** Frenkel defect, **(f)** Schottky defect

distinguished here depending on the way how these structures are formed:

- An ordered *superlattice structure* with the same basic crystal structure type may form upon cooling a disordered solid solution. The following example elucidates this scenario: a Cu alloy containing 25 at. % Au has a simple fcc crystal structure temperatures. Upon cooling it transforms into an ordered fcc $L1_2$ crystal structure of type Cu_3Au, where the gold atoms occupy the corner lattice points of the unit cell and the copper atoms sit on the faces of the cube.
- An intermetallic compound made up by two or more elements, producing a new phase with its own composition, crystal structure, and properties substantially different from those of its constituents, namely higher hardness, strength, and melting point but almost always at the expense of a lack of ductility. A recent, technically relevant example are the titanium aluminides (Fig. 3.6) considered for a variety of high-temperature applications, such as automotive valves and turbocharger wheels, and turbine blades and vanes in aerospace engines, as structural materials, mainly because of their attractive combination of high melting point and strength together with low density.

A final point should be noted here: many superlattice structures and intermetallic compounds have a range of compositions (*stoichiometry range*) in which they appear in the same crystal structure as compared with the fixed stoichiometric composition. This is, e.g., the case for both of the TiAl-based intermetallic compounds shown in Fig. 3.6, thus, making alloy property improvement and fine-tuning through addition of further alloying elements more feasible. This *nonstoichiometry* can lead to partial disordering of the atomic arrangement within the unit cell.

3.1.2 Microstructure

The crystal structure introduced in the previous section describes the ideal arrangement of atoms within a solid material and, hence determines several *intrinsic* material properties (e.g., elastic stiffness and compliance constants). However, lattice imperfections, which destroy the infinite extension of the periodic atomic structure, are decisive for many *extrinsic* properties (e.g., the mechanical properties discussed in Sect. 3.2.1). As a result, a real structure containing crystalline areas and a variety of lattice imperfections is called *microstructure*. Lattice imperfections are created and can be controlled during the processing and manufacturing of materials. They can be classified through their dimensionality as follows.

Point Defects

Figure 3.7 provides an overview of the various types of point defects (of zero dimensionality) in materials. Foreign atoms, either solved substitutionally on regular lattice sites (Fig. 3.7c,d) or interstitially (Fig. 3.7b), have already been treated in the previous section. If a lattice site is without an atom, we get a *vacancy* in the

lattice (Fig. 3.7a). From thermodynamics calculations we know that a certain (small) number of vacancies exists in thermal equilibrium. A more detailed treatment [3.11] yields

$$x_V^e = \exp\frac{-\Delta G_V}{RT} = \exp\frac{-(\Delta H_V - T\Delta S_V)}{RT}\,, \quad (3.1)$$

where x_V^e is the equilibrium concentration of vacancies, which follows an Arrhenius-type increase with increasing temperature. At the melting point of the material $x_V^e = 10^{-4}$–10^{-3}, whereas at absolute zero, i. e., 0 K, $x_V^e \equiv 0$. From the above it becomes obvious that vacancies will very likely be introduced into materials upon solidification from the melt. Figure 3.7b–d represents the various types of alloying elements in solid solution which were already treated in the previous section. Two special types of point defects are displayed in Fig. 3.7e and f: *Frenkel defects*, consisting of pairs of vacancy and interstitial atom, can be created upon neutron irradiation (Fig. 3.7e), while *Schottky defects* exist only in ionic crystals as pairs of vacancies formed in the cationic and anionic partial lattice, respectively (Fig. 3.7f).

Solid Solution Strengthening. It has been well known for a long time that introduction of alloying elements into a pure metal increases its hardness. When the difference in size and electronegativity of the alloying element is less than a critical value, the alloying element forms a solution with the matrix. The alloy atoms occupy either lattice sites, leading to a *substitutional solid solution*, or the interstitial voids in the lattice, i. e., *interstitial solid solutions*. Carbon, nitrogen, oxygen, hydrogen, and boron are elements that commonly occupy interstitial sites. Solute atoms interact with dislocations in a number of ways:

- *Par-elastic interaction* due to overlapping strain energy of solute atom and dislocation core. The interaction energy is directly proportional to the size difference between the solute and matrix atoms.
- *Modulus interaction* due to a local change of modulus, thereby affecting elastic energy of the dislocation. This is also called *Di-elastic interaction* in the literature [3.10]. The interaction energy is directly proportional to the difference in shear modulus between the solute and matrix atoms.
- *Stacking-fault interaction* or *Suzuki hardening* due to preferential segregation of solutes to the stacking fault of extended dislocations.
- *Electrical interactions* due to localization of electron cloud, leading to interaction with dislocations with electrical dipoles. This effect is usually smaller than the above mechanisms.

All these interactions require extra energy to be expended to overcome the solute atom, requiring higher stresses for dislocation motion and, hence, give rise to *solid solution hardening*. It may be pointed out that the presence of vacancies, introduced due to rapid quenching or high-energy radiations, also leads to considerable strengthening by some of the above mechanisms. For further details see [3.10].

Dislocations

Independently from each other Orowan, Polanyi, and Taylor introduced the term *dislocations* in 1934 in order to explain the observed strength values and the plastic deformability of materials on a theoretical basis. It should be emphasized here that the existence of these one-dimensional lattice imperfections was proven experimentally more than a decade later with the advent of the first transmission electron microscopes in materials science. To motivate why dislocations are essential in explaining the deformation behavior of materials we first consider the theoretical (shear) strength of materials shown in Fig. 3.8, in which A and B represent sites of stable equilibrium for the atoms within the hexagonal plane. Then a is the displacement required to shift any atom to another stable position, hence, the potential connected with a displacement x of any atom is

$$U(x) = U_0\left(1 - \cos\frac{2\pi x}{a}\right). \quad (3.2)$$

From (3.2) one can obtain the necessary force by differentiating with respect to x

$$F(x) = \frac{\partial U(x)}{\partial x} = U_0\frac{2\pi}{a}\sin\frac{2\pi x}{a}\,. \quad (3.3)$$

The shear stress is simply $\tau = F(x)/A$. For $x \ll a$ and since the sin function is point-symmetric about the origin, one gets from (3.3)

$$\tau \approx \tau_{max}\frac{2\pi x}{a} \quad (3.4)$$

with $\tau_{max} = U_0(2\pi x/aA)$. On the other hand, Hooke's law for simple shear is

$$\tau = G\gamma = G\frac{x}{d}\,. \quad (3.5)$$

Setting (3.4) and (3.5) equal, one obtains for the maximum shear stress

$$\tau_{max} = \frac{G\,a}{2\pi d} = \frac{G\,a}{2\pi\sqrt{3}} \approx \frac{G}{10}\,. \quad (3.6)$$

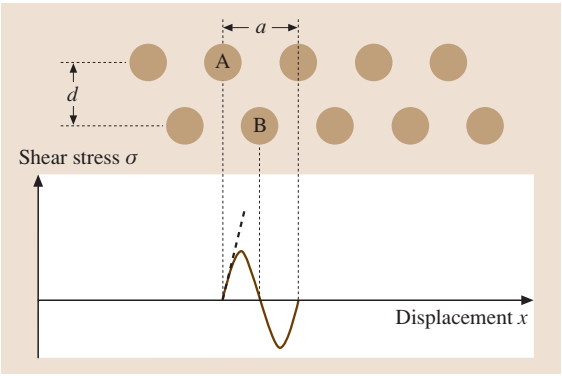


Fig. 3.8 The theoretical strength of crystalline materials: the shear displacement x between two neighboring lattice planes of interplanar spacing d causes a sinusoidal shear stress fluctuation σ. A and B are regular lattice sites, respectively, a is the lattice constant within the hexagonal plane

Thus, the theoretical shear strength τ_{max} is only dependent on the elastic properties of a material, i. e., the shear modulus G. As an example, for pure copper with $G \approx 45\,\text{GPa}$ one gets $\tau_{max} = 4500\,\text{MPa}$. However, in experiments one observes for pure copper single crystals that plastic shearing occurs already at shear stresses below 10 MPa [3.10] (Fig. 3.12). This obvious discrepancy by about three orders of magnitude can be rationalized only when assuming the existence of dislocations which enables plastic shearing by moving this dislocation in a direction $\boldsymbol{r}$ perpendicular to its line (the dislocation line is represented by the vector $\boldsymbol{s}$) through the lattice.

We can identify two basic types of dislocations:

1. *Screw dislocations*, which can be illustrated by cutting halfway through a perfect crystal (Fig. 3.9a) and subsequently skewing the crystal one atomic spacing (Fig. 3.9b,c). If one follows a crystallographic plane one revolution around the axis on which the crystal was skewed, starting at point x and moving an equal number of atomic spacings in each of the four planar directions one ends up one lattice site below the starting point (point y). The vector required to close the loop is called the *Burgers vector* $\boldsymbol{b}$ and represents the unit of plastic shearing of the crystal through the moving of the dislocation line (from

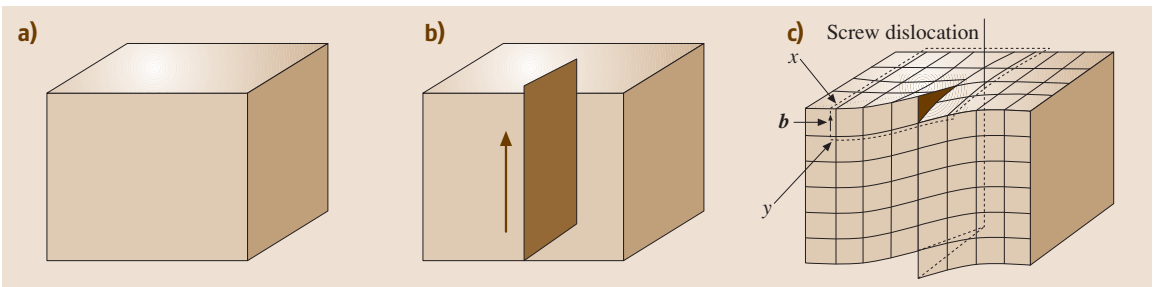


Fig. 3.9a–c A perfect crystal (a) is cut and (b) and (c) sheared by one atomic spacing. The line along which shearing is carried out is a screw dislocation with its Burgers vector $\boldsymbol{b}$ closing a loop of equal atomic spacings around itself

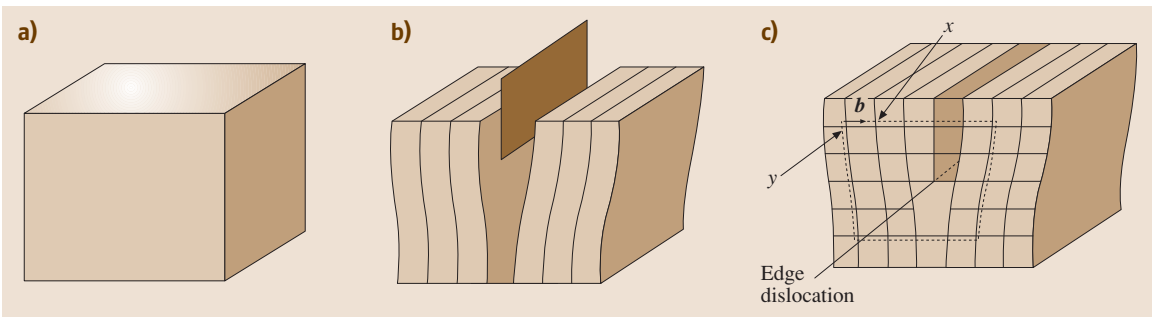


Fig. 3.10a–c A perfect crystal (a) is cut and an extra (half) plane of atoms is inserted (b). The bottom edge of the extra plane is an edge dislocation with its Burgers vector $\boldsymbol{b}$ closing a loop of equal atomic spacings around itself

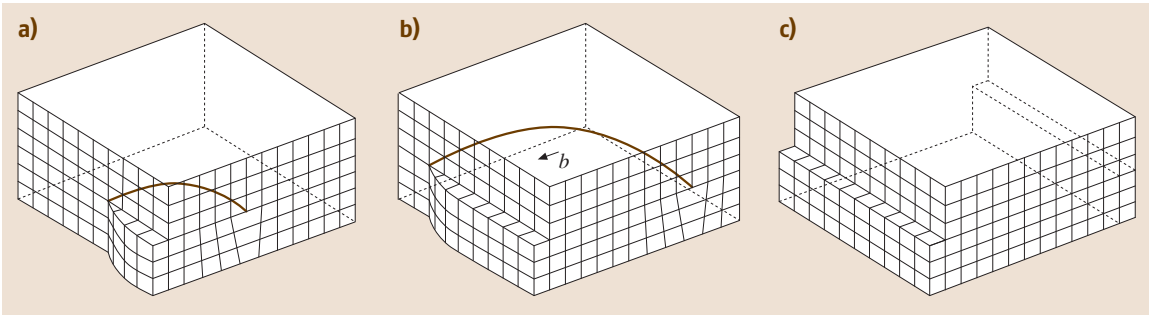


Fig. 3.11a–c Movement of a mixed dislocation through a simple cubic lattice: **(a)** and **(b)** the dislocation character is purely screw-type on the left side and purely edge-type on the right side of the crystal; **(c)** after completion of the movement the dislocation has disappeared and has left a slip step of height b

the front to the back in Fig. 3.9c). Hence, in screw dislocations $\boldsymbol{b} \parallel \boldsymbol{s} \perp \boldsymbol{r}$.

2. *Edge dislocations*, which can be illustrated by slicing halfway through a perfect crystal (Fig. 3.10a), tearing the crystal apart and inserting an extra (half) plane of atoms into the cut (Fig. 3.10b). The bottom edge of this inserted plane represents the edge dislocation (Fig. 3.10c). A clockwise loop, starting at point x and going an equal number of lattice sites into each direction within the plane finishes at point y, hence leaving the required Burgers vector $\boldsymbol{b}$ to close the loop in Fig. 3.10c. For edge dislocations $\boldsymbol{r} \parallel \boldsymbol{b} \perp \boldsymbol{s}$.

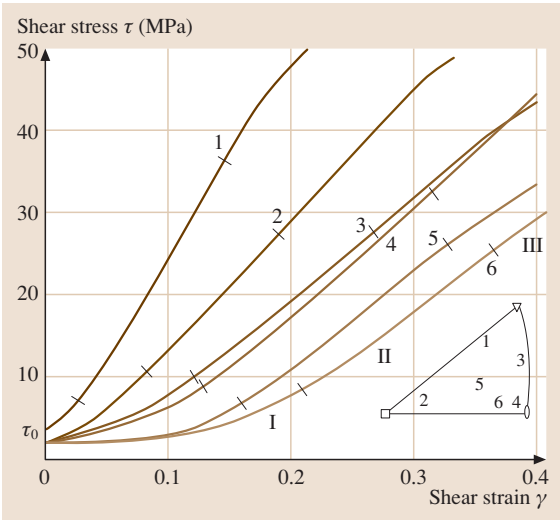


Fig. 3.12 Shear stress versus shear strain curves for differently oriented copper single crystals at room temperature. Samples 1–6 are oriented within the standard orientation triangle as depicted

The more general case of a dislocation with *mixed* character when $\boldsymbol{b}$ is neither parallel nor perpendicular to $\boldsymbol{s}$ is plotted in Fig. 3.11. Likewise, one notes from Fig. 3.11 that pure screw or edge configurations can be considered the extremal cases of the general mixed configuration for either $\boldsymbol{b} \parallel \boldsymbol{s}$ or $\boldsymbol{b} \perp \boldsymbol{s}$.

From the above the following important conclusions can be drawn:

- The dislocation line is the border between the undeformed and the slipped area of the crystal plane.
- Dislocations cannot end within a crystal: they either have to form a loop or penetrate an internal (e.g., grain or phase boundaries, see below) or external surface.
- Dislocations exhibit a long-range three-dimensional stress field within the crystal lattice, allowing them to interact with all kinds of other lattice imperfections. This feature is the basis for the various approaches of strengthening materials in metallurgy.
- Dislocations are the carrier of plastic (i. e., irreversible) deformation in crystalline materials.
- The slip system of a dislocation in crystalline materials is composed of:

1. A close packed slip plane
2. A close packed slip direction which is identical with $\boldsymbol{b}$

The Burgers vector must be contained within the slip plane. Both, slip plane and slip direction depend on the lattice structure of the material.

Similarly as for the grain boundaries explained below, dislocations are created in crystals during solidification due to thermal stresses arising from the density mismatch between the liquid phase and the solid phases.

Hence, the creation of a dislocation within a crystal increases the (total) free energy of the system. A satisfactory approximation for the dislocation line energy is

$$E_\rho = \frac{1}{2} G b^2 \,. \tag{3.7}$$

Equation (3.7) reveals that the dislocation line energy is an *intrinsic* materials property, depending only on crystal lattice parameters. Experimentally, dislocations can be detected either via optical microscopy as *etch pits* stemming from the penetration points of the dislocation lines through the crystal surface or via transmission electron microscopy (TEM), where the slight lattice distortions caused by the dislocations gives rise to a contrast visible in the TEM; see [3.12] for details. For explaining the strength and deformability of materials it is useful to introduce the dislocation density for the number of dislocation per unit volume as

$$\rho = \frac{\text{dislocation line length}}{\text{unit volume}}$$

$$= \frac{\text{number of dislocations}}{\text{unit area}} \,. \tag{3.8}$$

As an example for well-annealed metals $\rho = 10^9 - 10^{11}\,\mathrm{m}^{-2}$, whereas for ceramics and semiconductors $\rho = 10^4 - 10^{10}\,\mathrm{m}^{-2}$.

Strain Hardening. Strain hardening is caused by interaction of dislocations with each other. During plastic deformation, e.g., cold working, the number of dislocations increases with increasing strain to values of $\rho = 10^{14} - 10^{16}\,\mathrm{m}^{-2}$. Similarly, an increase of dislocation density is connected with tensile straining. An example for the stress–strain behavior of Cu single crystals with different crystallographic orientations with respect to the loading axis is depicted in Fig. 3.12. The *flow curve* comprises three stages:

1. Stage I or *easy glide*
2. Stage II or *dislocation pile-ups*
3. Stage III or *dynamic recovery*

Stage I is observed only in well-annealed crystals oriented such that only one slip system is operative, cf. curves 5 and 6 in Fig. 3.12. Stage II dominates the flow curve of most engineering polycrystalline alloys and measurements over a wide range show that

$$\sigma_\rho = \sigma_\mathrm{i} + \alpha G b \sqrt{\rho} \,, \tag{3.9}$$

where α is a numeric constant (typically ≈ 0.3). Equation (3.9) is also referred to as Taylor's relation and originates from the interaction of the stress fields of dislocations (long-range interaction) or from dislocations cutting due to intersecting slip systems. The strain hardening rate during stage II is fairly insensitive to temperature and/or impurities. On the other hand, the region of dominance of stage III, *dynamic recovery*, is strongly temperature dependent. For further details see, e.g., [3.13].

Grain Boundaries

Grain boundaries are two-dimensional lattice faults. Like vacancies and dislocations they are created during manufacturing of materials, as can be exemplified with metallic materials upon solidification from the amorphous (or disordered) melt (Fig. 3.13a). When decreasing the temperature of the melt below the liquidus temperature (Sect. 3.1.2) nuclei that have the crystal structure of the solid are formed within the liquid and

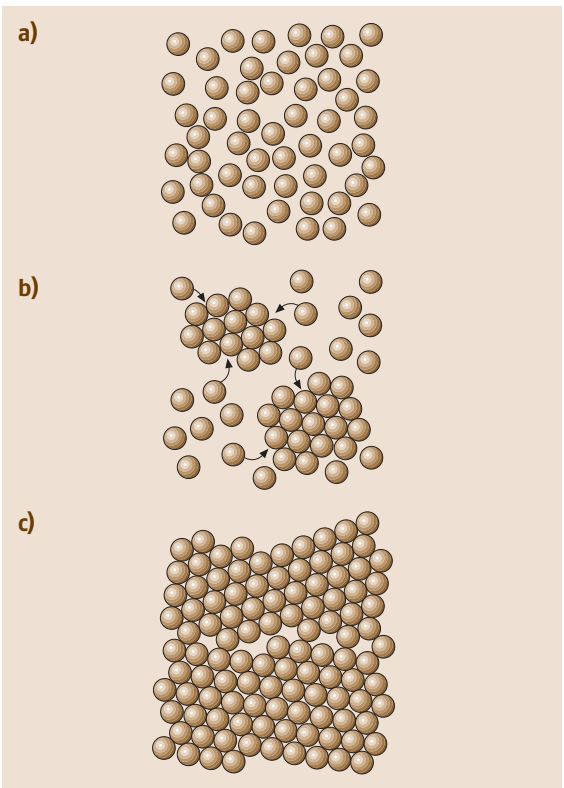


Fig. 3.13a–c The development of grains and grain boundaries upon solidification from the melt: **(a)** amorphous melt, **(b)** two crystals begin to nucleate within the liquid melt, **(c)** a grain boundary has been created between the two crystallites of different crystallographic orientation

grow with continuing time and cooling (Fig. 3.13b). When T falls below the solidus temperature the material has solidified completely and adjacent grains with the same crystal lattice but different crystallographic orientation (Fig. 3.13c) do not fit perfectly to each other. The narrow zone where the atoms are not properly spaced is called a *grain boundary*. Typically, $d_{gb} = 0.5$ nm is a good estimate for the grain boundary thickness.

Fine Grain Strengthening. One important method of controlling the properties of a material is to adjust the grain size. By reducing the grain size, the number of grains and hence the fraction of grain boundaries is increased. If mechanical properties are concerned, any dislocation that moves within a grain is stopped when it encounters a grain boundary. The mean free path of dislocations is, thus, limited by the grain size and the strength of the metal is increased. The famous Hall–Petch relation relates the grain size to the yield strength at room temperature [3.15, 16]

$$\sigma_y = \sigma_0 + K_y d_{gb}^{-1/2} , \qquad (3.10)$$

where σ_y is the yield strength (Sect. 3.2.1), $d_{gb}^{-1/2}$ is the inverse square root of grain size, preferentially measured as mean intercept length (3.44) in Sect. 3.2.3, and σ_0 and K_y are constants for material (σ_0 is often related to the *Peierls stress* in crystallographic slip [3.10]). Figure 3.14 shows the relationship according to (3.10) for some steels.

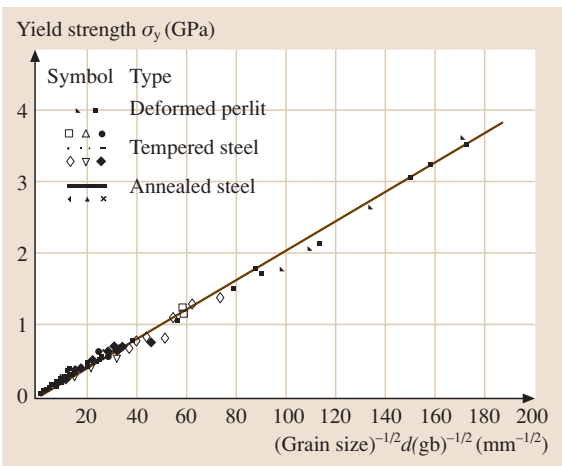


Fig. 3.14 Dependence of yield strength σ_y on the inverse root of the grain size, $d_{gb}^{-1/2}$, in steel (after [3.14])

Optical or scanning electron microscopy (SEM) can be used to reveal microstructural features such as grain boundaries (Sect. 3.2.3) and to assess the grain size of materials quantitatively.

Dispersoids and Precipitates

When the solubility of a material is exceeded for any alloying element, a second phase forms in the volume and a two-phase alloy is produced. Hence, second-phase particles such as *dispersoids* or *precipitates* are classified as three-dimensional lattice imperfections. The continuous phase that surrounds the particles and is usually present in a larger fraction, is called the *matrix*. The boundary between the two phases is an internal surface or *interface* at which, as for the grain boundaries in the previous section, the atomic arrangement is not perfect. Again, this boundary impedes the slip of dislocations and, thus, strengthens the material.

Dispersion and Precipitation Strengthening. Two types of second-phase particles can be distinguished. In *dispersion strengthening*, hard particles are introduced into the matrix using powder metallurgical techniques or through solid reactions. These particles are in essence insoluble in and *incoherent* (Fig. 3.15a) with the matrix. *Precipitation strengthening* or *age hardening* is produced through a series of heat treatments that exploit the decrease in solubility of a given solute with decreasing temperature. This requirement for elevated temperature solubility places a limitation on the number of useful precipitate-strengthened alloy systems. Due to their insolubility within the metallic matrix, dispersion-hardened alloys are stable up to temperatures relatively close to the melting point of the matrix, in contrast to precipitate-strengthened alloys which are degraded upon prolonged exposure to high temperature due to *precipitate coarsening*, i. e., *Ostwald ripening*. Nevertheless, precipitation hardening due to very small *coherent* particles (2–10 nm, Fig. 3.15b) is a very efficient strengthening mechanism. Fine particles can act as barriers to dislocation motion either by requiring the dislocations to shear them or by acting as strong impenetrable particles, forcing dislocations to bypass them. When the particles are small and/or soft they get sheared and the following mechanisms contribute to the strength increment:

- *Coherency strain*, arising from the strain field resulting from mismatch between particle and the matrix.
- *Stacking fault energy* variation between matrix and particles, which leads to local variation in the stacking fault width.

- If the precipitates have an *ordered structure*, as in many Ni-based superalloys (Sect. 3.7.5), motion of dislocation through them introduces antiphase boundaries (APB).
- A difference in *local modulus* of the particle and matrix alters the energy of the dislocation passing through the particle. Likewise, voids counterintuitively can contribute significantly to strength through this mechanism.
- As dislocations pass through a precipitate a step, which is one Burgers vector high, is produced. This raises the particle–matrix *interfacial energy*, contributing to strengthening.
- Finally, there is a strengthening contribution due to difference in lattice friction stress or Peierls stress in particle and matrix.

For larger (incoherent) precipitates or dispersion strengthened alloys, Orowan proposed a mechanism for particle overcoming, as illustrated in Fig. 3.16. In essence, the strength increase is related to the increase of dislocation line length when the dislocation deposits a loop around the particle after bypassing it (Fig. 3.16d). The critical situation for bypassing is reached when the dislocation line possesses its maximum curvature around the particle, i. e., the half-circle. According to Fig. 3.17 one obtains for the critical configuration for dislocation bypassing ($\theta \to 0°$)

$$F = \frac{\tau b}{l} = 2E_\rho = Gb^2 \,. \tag{3.11}$$

Solving (3.11) for τ one obtains the strength increase due to Orowan bypassing

$$\tau_{OR} = \frac{Gb}{l} \,. \tag{3.12}$$

Since the particle spacing l is not directly extractable from microstructure analysis (Sect. 3.2) the following more useful expression has been established for the strength increment from Orowan bypassing which also takes into account the conversion of the shear stress τ into a normal stress σ via the Taylor factor $M = \sigma/\tau$

$$\sigma_{\mathrm{OR}} = 0.8MGb\frac{\sqrt{f}}{r} \,, \tag{3.13}$$

where f is the volume fraction of particles present in the material and r is the (directly measurable) particle radius. For a detailed review on the various types of particle strengthening sketched above see [3.17].

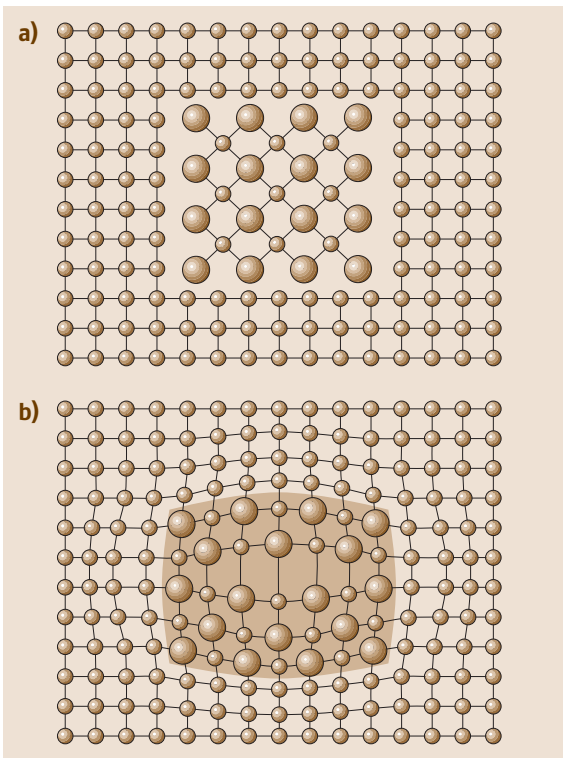


Fig. 3.15 **(a)** An incoherent second-phase particle has no crystallographic relationship with the structure of the surrounding matrix; **(b)** coherent precipitates show a definite crystallographic relationship with the matrix

3.1.3 Atomic Movement in Materials

The movement of atoms (or molecules) within materials (or liquids) is called *diffusion*. It is emphasized here

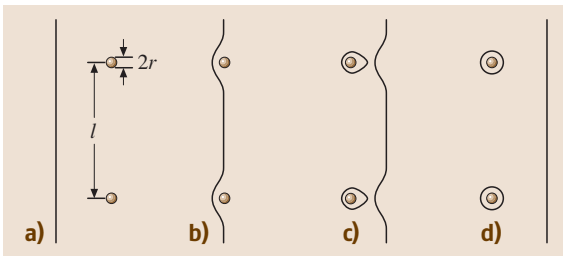


Fig. 3.16a–d Scheme of dislocation bypassing of fine nonshearable particles by Orowan bowing: driven by shear stress τ, a dislocation approaches an array of (two) particles with radius r separated by a distance l **(a)**. It bows out between the particles **(b)**, until it deposits loops around them **(c)**. After by-passing the dislocation line remains unchanged **(d)**

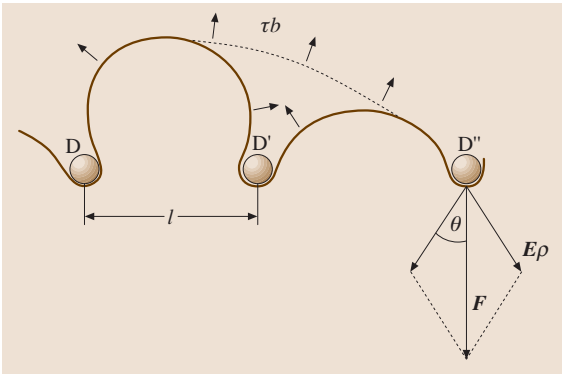


Fig. 3.17 Bowing out of a dislocation line between an array of impenetrable incoherent particles (designated D, D′, D″) with spacing l due to a shear force (per unit length of dislocation) τb. The obstacle force acting against bypassing is F

that diffusion can take place without a superimposed mechanical stress just by random movement of atoms, provided that enough thermal energy is introduced into the system. Vice versa, as for the concentration of vacancies, (3.1) in Sect. 3.1.2, atoms in crystalline solids are at rest only at absolute zero. The driving force behind this movement is usually a gradient in concentration of atoms according to

$$j_D = -D\,\mathbf{grad}\,c = -D\frac{\partial c}{\partial x}\,, \qquad (3.14)$$

where the latter fraction is the *concentration gradient* in the simplified case for one-dimensional atom movement. Equation (3.14) is called *Fick's first law* and is visualized in Fig. 3.18. j_D is the flux of atoms through a plane of unit area per unit time, and D is the *diffusion coefficient* (m^2/s).

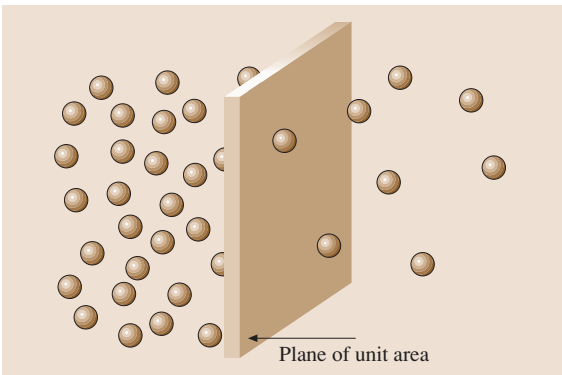


Fig. 3.18 The flux j_D during diffusion is defined as the number of atoms passing through a plane of unit area per unit time

Consider now a sheet of nickel and copper bonded to each other (Fig. 3.19a) the steep concentration gradient at the contact surface will cause continuous mutual diffusion of both atomic species to either side (Fig. 3.19b) until eventually an equilibrium concentration is attained (Fig. 3.19c). The kinetics, i. e., the velocity of this interdiffusion process is triggered by the absolute temperature T according to an Arrhenius law

$$D = D_0 \exp\left(-\frac{Q}{RT}\right). \qquad (3.15)$$

D_0 is a constant prefactor and – in essence – an intrinsic material parameter, $R = 8.314\,\mathrm{J mol^{-1} K^{-1}}$ is the gas constant and Q is the activation energy required for the atom to carry out a single jump event. In crystalline solids, as pointed out in Sect. 3.1.2, two mechanisms of atom movement are conceivable depending on the size ratio δ of solute and matrix atom (Sect. 3.1.1):

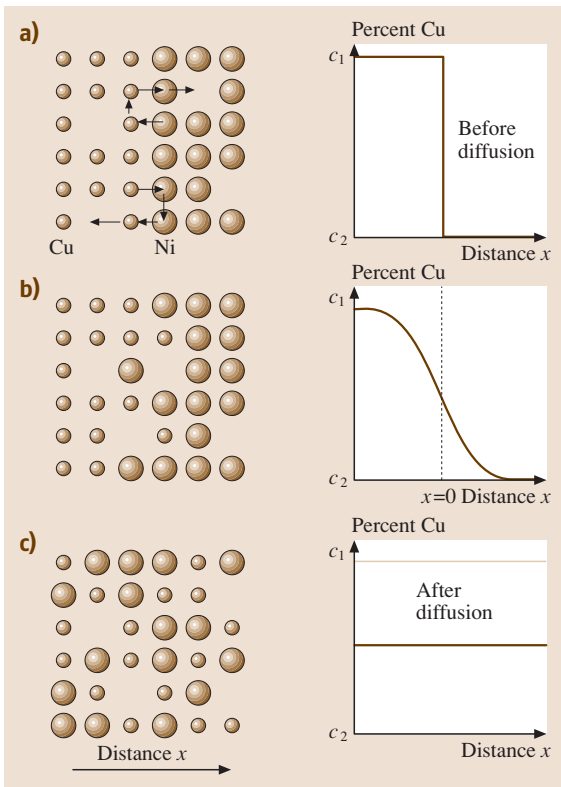


Fig. 3.19a–c Mutual diffusion of copper and nickel atoms into each other through a vacancy mechanism: **(a)** $t = 0$, **(b)** $t > 0$, intermediate time, **(c)** $t \to \infty$

- $\delta < 0.15$: this is the case for self-diffusion and diffusion of substitutionally solved atoms and requires the presence of a vacancy next to the lattice site of the moving atom (see the upper part of Fig. 3.20). After the jump the atom has created a new vacancy at the original lattice site, hence, one observes a countercurrent flow of atoms and vacancies. Consequently, this mechanism is also called *vacancy diffusion*. Since this diffusion mechanism requires the presence of vacancies, its activation energy Q_v is composed of two terms, namely one for the formation of vacancies, Q_f, and one for their migration, Q_m, hence $Q_v = Q_f + Q_m$ replaces Q in (3.15) for self-diffusion or vacancy diffusion.
- $\delta > 0.15$: this is the case for (small) interstitial atoms moving from one interstitial site to another. No vacancies are required for this mechanism and the activation energy for *interstitial diffusion*, Q_i, accounts for the migration of interstitials and is therefore smaller than its counterpart for vacancy diffusion Q_v (Fig. 3.20). For interstitial diffusion Q_i substitutes for Q in (3.15).

Examples for the temperature dependence of the diffusion coefficient according to (3.15) are given in Fig. 3.21 in the form of an Arrhenius-type plot of ln D versus the reciprocal temperature $1/T$. One notes the following characteristic features:

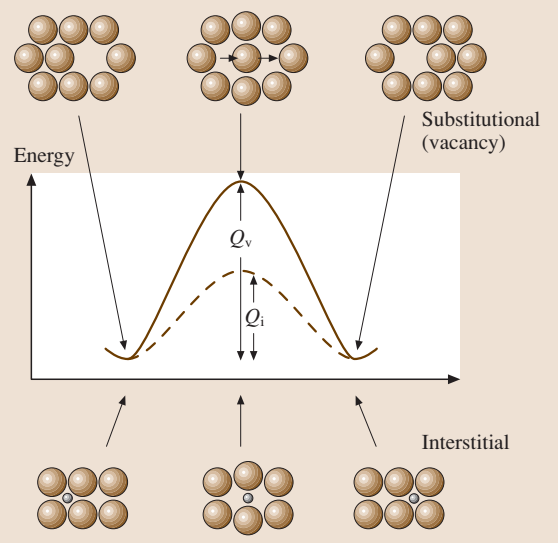


Fig. 3.20 Visualization of diffusion mechanisms in crystalline solids: *top* vacancy diffusion with activation energy Q_v, *below* interstitial diffusion with activation energy Q_i

1. The slope of the curves is a measure for the activation energy Q, with a steeper slope indicating a higher value of Q.
2. As anticipated from Fig. 3.20, interstitial diffusion is considerably faster than vacancy diffusion, cf. the curves for Fe self-diffusion with those of carbon and hydrogen diffusion in iron.
3. The lower packing density (PD) of the bcc crystal structure as compared with the fcc or hcp structure (Table 3.2) gives rise to a higher diffusivity D.

This holds for Fe as well as for Ti, which undergoes an allotropic transformation from hcp α-Ti to bcc β-Ti at 882 °C.

The scenario depicted in Fig. 3.19 is typical for many engineering applications of diffusional processes such as heat treatment of materials for equilibrating concentration inhomogeneities stemming from alloy solidification, joining operations. Figure 3.19 can be considered as an exemplification of diffusional bonding, or consolidation of metallic and/or ceramic powder particles through solid-state sintering (*powder metallurgy*). Fick's first law (3.14) is unable to describe the local and temporal distribution of atoms during different time stages of diffusion, e.g., the concentration profile shown in Fig. 3.19b. However, as the number of atomic species in the system remains constant

$$\frac{\partial c}{\partial t} + \operatorname{div} j_D = 0 \tag{3.16}$$

and with (3.14) one obtains finally after some manipulations (see [3.11] for a detailed derivation) Fick's second law (in its one-dimensional form)

$$\frac{\partial c}{\partial t} = D\frac{\partial^2 c}{\partial x^2}\,, \tag{3.17}$$

whose solution depends on the specific boundary conditions of the diffusion problem. For the scenario depicted in Fig. 3.19 (*diffusion bonding* of two semi-infinite rods of different metals) one obtains

$$\frac{c(x,t) - c_1}{c_2 - c_1} = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x}{2\sqrt{Dt}}\right)\right] \tag{3.18}$$

with erf(ξ) being the *error function* (also called the *Gaussian probability integral*), which can be solved only numerically. Note that the error function is point-symmetric with respect to the origin and $-1 \le \operatorname{erf}(\xi) \le 1$ for $-\infty \le \xi \le \infty$. Equation (3.18) reveals that, for obtaining a certain given concentration c_0 at depth x_0 both, temperature (via D) and time t can be varied independently to design an optimum heat treat-

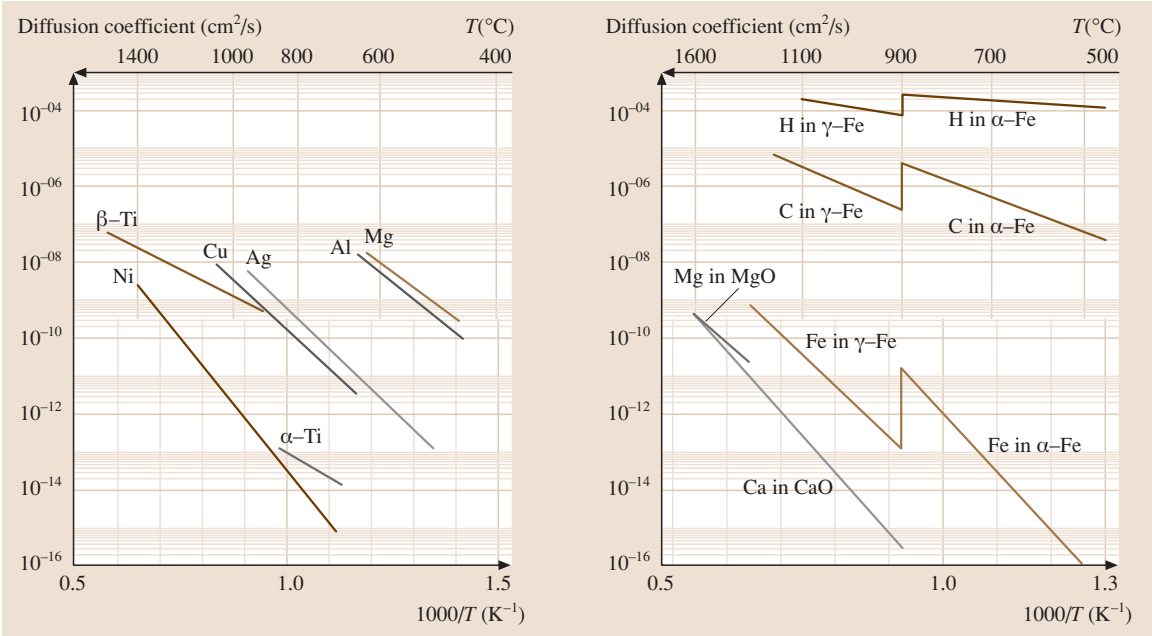


Fig. 3.21 Temperature dependence of the diffusion coefficient D in various metals and oxide ceramics. The slope of the ln D versus $1/T$ plot represents the activation energy for diffusion Q

ment in view of the availability of appropriate furnaces as well as cost of operation. The most prominent application of diffusional problems is the carburization of steels, for which solutions similar to (3.18) are available in literature [3.11].

3.1.4 Transformation into Solid State

In Sect. 3.1.2 we have already noted that the solidification process of materials critically influences the microstructure in general, and the amount of dislocations and grain boundaries in particular. While physical (PVD) and chemical (CVD) vapor deposition techniques on (cold) surfaces become increasingly important in thin-film technologies, the main and most important phase transformation in metallic materials, semiconductors, and glasses in terms of both mass production and annual turnover is still by far solidification from the melt (*cast metallurgy*). This phase transformation can be quantitatively treated by applying the principles of thermodynamics. Hence, in this section we will introduce the basic thermodynamic concepts focusing on the behavior of pure materials, in other words a single-component *system*, a material that can exist as a mixture of one or more phases. A *phase* can then be defined as apportion of the system whose properties and composition are homogeneous and which is physically different from other parts of the system. The *components* are the different (chemical) elements which make up a system, and the composition of a phase or the system can be described by giving the relative amounts of each component.

Consequently, the subsequent sections show how solidification occurs in alloys and multiple-*phase* systems and the main species of binary phase diagrams are derived.

The reason why a transformation occurs at all is because the initial state of the material is unstable relative to the final state. This scenario can be expressed by thermodynamics principles (at constant temperature and pressure p) through the Gibbs free energy G of the system

$$G = H - TS\,, \tag{3.19}$$

where H is the enthalpy, i. e., the heat content arising (to a good approximation for condensed matter) from the total kinetic and potential energies of the atoms, and S is the entropy, i. e., a measure of the randomness of the system.

A system is in *equilibrium* when it is in its most stable state. This translates (3.21) for a closed system (of fixed mass and composition) at constant T and p

into

$$dG = 0\,. \tag{3.20}$$

From (3.19, 3.20) one can intuitively conclude that the state with the highest stability will be the one with the best compromise between a low value of H and a high entropy. Thus, at low temperatures, solid phases are most stable since they have the strongest atomic binding and, hence, lowest enthalpy. At higher temperatures, however, the $-TS$ term in (3.19) dominates over H and phases with increasingly larger degree of atom movement become stable: first liquids and then gases. This is elucidated in Fig. 3.22 where the intersection of the curves with the ordinate is a measure of the enthalpy of the respective phases and the slope of the curves represent the entropy.

Homogeneous Nucleation

Undercooling a liquid below its equilibrium temperature T_m yields a driving force for solidification $\Delta G = (G_S - G_L) < 0$, so one might expect the melt to solidify spontaneously. However, this is not the case and liquids can be supercooled by more than a hundred Kelvin below T_m without crystallization [3.5–8] when a nucleus of solid matter has to be formed within the homogeneous liquid. The change in free energy of the system ΔG_{cryst} when producing a solid sphere of radius r within the liquid for a given undercooling ΔT consists of two terms

$$\Delta G_{cryst} = -\frac{4}{3} r^3 \pi \Delta G + 4r^2 \pi \gamma_{SL}, \tag{3.21}$$

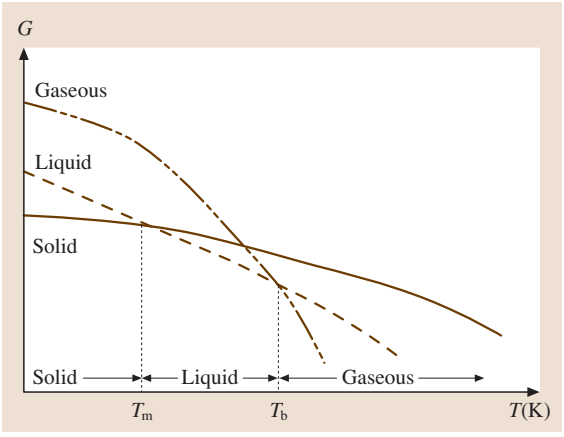


Fig. 3.22 Differences in molar free energy ΔG between solids, liquids, and gases in a single-component system (pure material). T_m and T_b denote the melting and boiling point, respectively

of which the first term, ΔG_V, is the gain in energy due to the formation of a spherical volume of crystalline solid and the second term, ΔG_O, is the expenditure of energy due to the formation of the interface of specific interfacial energy γ_{SL} between the liquid and solid phases. The individual terms ΔG_V and ΔG_O are plotted together with the sum curve ΔG_{cryst} as a function of crystal radius r in Fig. 3.23.

The critical radius $r*$ of a stable nucleus of crystalline solid is reached when further growth of the nucleus leads to a gain in ΔG_{cryst}. Mathematically, this is obtained for the first derivative of ΔG_{cryst} in (3.21) with respect to r

$$\frac{\partial \Delta G_{cryst}}{\partial r} = 0 = -4r^{*2}\pi \Delta G + 8r^* \pi \gamma_{SL}\,. \tag{3.22}$$

Solving (3.22) yields for the critical radius

$$r^* = \frac{2\gamma_{SL}}{\Delta G} \tag{3.23}$$

and for the excess free energy

$$\Delta G^* = \frac{16\pi \gamma_{SL}^3}{3(\Delta G)^2}\,. \tag{3.24}$$

Since ΔG is proportional to the undercooling ΔT, (3.23, 3.24) straightforwardly demonstrate that small undercoolings require a large amount of ΔG^* and the

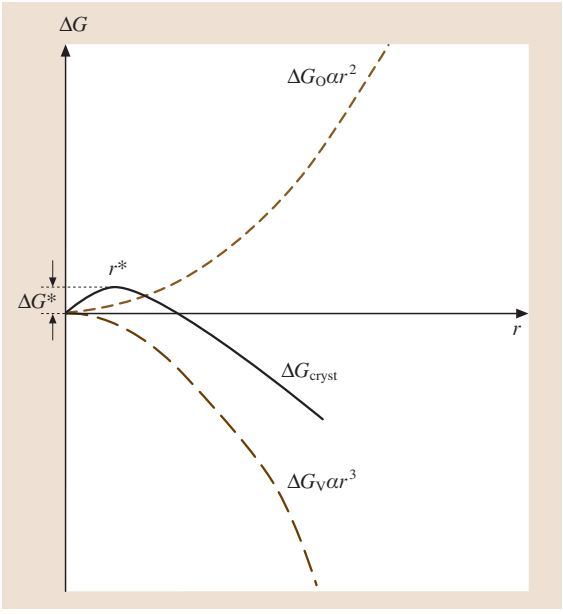


Fig. 3.23 Free-energy change associated with homogeneous nucleation of a sphere of radius r

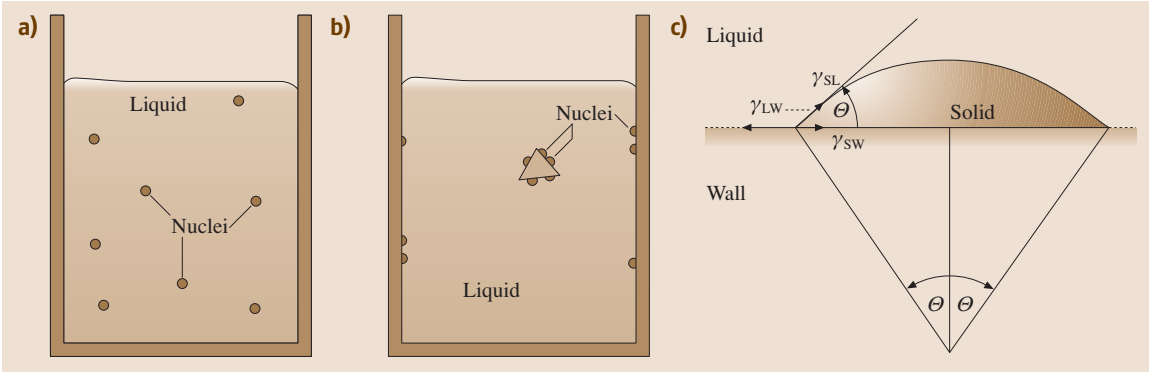


Fig. 3.24a–c Homogeneous (a) versus heterogeneous nucleation ((b), (c)): a spherical cap of a solid needs fewer atoms to become a stable nucleus than a sphere, see text

formation of large cluster of atoms for a stable nucleus. Vice versa, for large undercoolings, nuclei with a small critical radius are already stable within the melt and can grow.

Heterogeneous Nucleation

From (3.26) for ΔG^* it is obvious that nucleation can become easier if the interfacial energy term is reduced. This is effectively achieved if the nuclei form in contact with the mould wall (or likewise, impurities within the melt). Consider a solid embryo in contact with a perfectly flat mould wall (Fig. 3.26b,c). For a given volume of solid the total interfacial energy of the system can be minimized if the condition

$$\gamma_{LW} = \gamma_{SW} + \gamma_{SL} \cos \Theta \qquad (3.25)$$

is fulfilled in the plane of the mould wall. The embryo has the shape of a spherical cap with radius r and a wetting angle Θ. *Porter* and *Easterling* [3.11] compared this situation with that ofa sphere with same radius r which undergoes homogeneous nucleation (Fig. 3.24a) and derived the following relation for the energy barrier for heterogeneous nucleation

$$\Delta G^*_{het} = \frac{16\pi\gamma_{SL}^3}{3(\Delta G)^2} S(\Theta) , \qquad (3.26)$$

where

$$S(\Theta) = \frac{(2+\cos\Theta)(1-\cos\Theta)^2}{4} . \qquad (3.27)$$

Except for the shape factor $S(\Theta)$, (3.26) is the same as the relation obtained for homogeneous nucleation (3.24). Since $S(\Theta) \leq 1$ and the critical radius $r*$ is unaffected, heterogeneous nucleation is always energetically favored over homogeneous nucleation and, thus, also the rate of heterogeneous nucleation becomes faster [3.11]. If, for example, $\Theta = 10°$, $S(\Theta) \approx 10^{-4}$ and the energy barrier for heterogeneous nucleation becomes dramatically smaller than that for homogeneous nucleation. Even for the upper limit $\Theta = 90°$ (half sphere), $S(\Theta) = 0.5$.

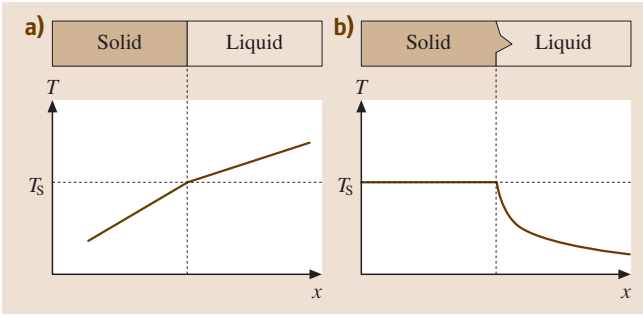


Fig. 3.25a,b Temperature distribution for solidification and the form of the solid–liquid interface when the heat is conducted through (a) the solid and (b) the liquid

Heat Flow and Interface Stability

Neglecting the effect of alloying, solidification is controlled by the rate at which the heat of crystallization is conducted away from the solid/liquid interface. Two options are conceivable:

1. If the solid grows with a planar interface into a superheated liquid (Fig. 3.25a), the heat flow away from the interface through the solid must balance that from the liquid plus the heat of crystallization generated at the interface. Then, a small branch of solid protruding into the liquid will arrive in a region of increased temperature. Consequently, more heat will be conducted into the protruding solid and less

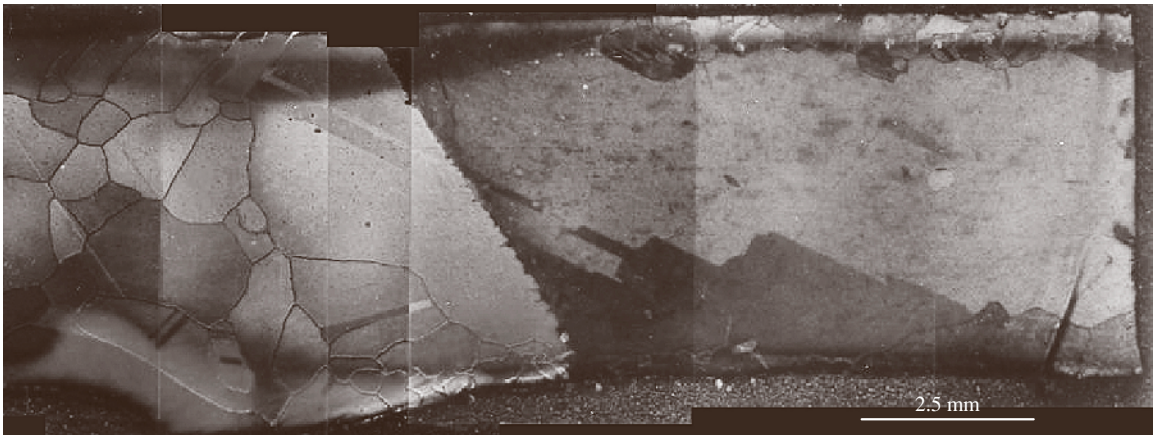


Fig. 3.26 Microstructure of a Ni−Cr20 alloy produced by zone melting

heat will be transported away such that the growth rate will slow down below that of the surrounding planar region and the protrusion will eventually disappear and the solid–liquid interface remains planar. Heat conduction through the solid, as depicted in Fig. 3.25a, can be promoted when solidification takes place from the cooler walls of the mould. This effect is applied technically by using a cold plate or a zone melter to apply a temperature gradient for producing microstructures consisting of coarse and elongated grains or even single crystals. An example for a Ni − Cr20 alloy is shown in Fig. 3.26: starting from the left side with coarse, but equiaxed grains the temperature gradient causes a single dominant grain to grow preferentially towards the right of the sample.

2. If the solid grows into a supercooled liquid (Fig. 3.25b), an eventual protrusion of the solid into the liquid is forced to grow more rapidly by the negative temperature gradient in the liquid because the heat is removed more efficiently from the tip of the protrusion than from the surrounding regions. Thus, a solid–liquid interface advancing into a supercooled liquid is inherently unstable.

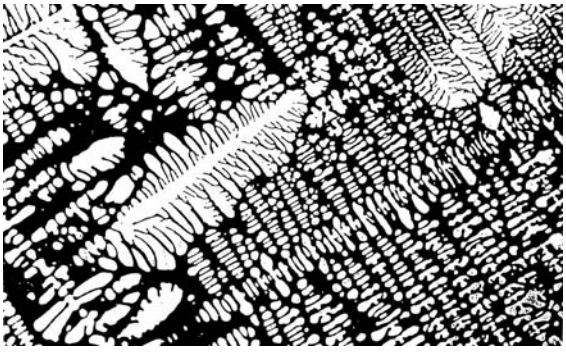


Fig. 3.27 Dendritic microstructure in a die-cast Al−Zn10−Si8−Mg alloy

In alloys the formation of dendrites is connected with compositional changes (or constitutional effects) between the solid and liquid phase, therefore dendrite formation is known to be caused by *constitutional supercooling*. An example for a dendritic microstructure in a cast Al alloy is shown in Fig. 3.27. For further details and a more quantitative treatment of these issues see [3.11].

3.1.5 Binary Phase Diagrams

In single-component systems all phases have the same composition, and equilibrium involves temperature T and pressure p as variables. Obviously, in alloys composition is also variable and understanding phase transformations requires an appreciation of how the Gibbs free energy of the respective phases involved depends on all these parameters. However, pressure p can be ruled out and treated as being constant in what follows since we consider only the liquid–solid transformation with both phases being essentially incompressible. Besides, to keep the physical model simple we restrict ourselves in the following to binary alloys, i. e., two-component systems.

Gibbs Free Energy of Binary Solutions

Assume that two components A and B can be mixed in any proportions (because they have the same crys-

tal structure; an example of this scenario would be the alloy system Cu – Ni). Then the Gibbs free energy of a homogeneous solid solution can be calculated in two steps: first, we bring together x_A mole of pure A and x_B mole of pure B, so the free energy of the system is simply the arithmetic mean

$$\bar{G}(x) = G_A x_A + G_B x_B , \tag{3.28}$$

with $x_A + x_B = 1$ for binary systems. For all alloy compositions $\bar{G}(x)$ lies on a straight line between G_A and G_B. The second step is now to let A and B mix in a random fashion. The free energy of the system will not remain constant during mixing and

$$G(x) = \bar{G} + \Delta G_M , \tag{3.29}$$

where $\Delta G_M = \Delta H_M - T\Delta S_M$ is the change in Gibbs free energy due to mixing. The simplest case of mixing is the *ideal solution*, in which one assumes that no preferences can be found for the different types of interatomic bonds between neighboring atoms. If preferred bondings are found, however, this situation is called a *regular solution* and describes more realistic scenarios of alloying; this case will not be treated here for simplicity. For ideal solutions $\Delta H_M = 0$ and $\Delta G_M = -T\Delta S_M$. From statistical thermodynamics [3.2,3] we know that entropy is quantitatively related to randomness by Boltzmann's equation and one obtains

$$\Delta G_M = -T\Delta S_M = RT(x_A \ln x_A + x_B \ln x_B) . \tag{3.30}$$

Since $\Delta S_M > 0$, i. e., there is an increase in entropy on mixing, ΔG_M is negative and the course of $G(x)$ according to (3.29) is shown schematically in Fig. 3.28. Note, that, as the temperature increases, G_A and G_B decrease and the free energy curve $G(x)$ assumes a greater curvature due to the increasing degree of randomness.

For any given mole fraction x_B the extrapolation of the tangent to G to both sides of the molar free energy diagram (Fig. 3.28) yields the chemical potentials μ_A and μ_B of the components A and B, respectively, which describe how the free energy changes when an infinitesimally small quantity of the atomic species i is added to the system (at constant T and p). Consequently

$$\mu_i = \left(\frac{\partial G}{\partial n_i}\right) . \tag{3.31}$$

Equilibrium in Heterogeneous Systems

Assume now that the components A and B do not have the same crystal structure. In this case, two free energy curves G_α and G_β have to be sketched, as shown in Fig. 3.29, and the stable forms of both structures are those with the lower free energy; thus, in thermodynamic equilibrium a homogeneous α solid solution is found for A-rich compositions and β is the stable phase for B-rich compositions. For alloy compositions near the crossover in the $G(x)$ curves (see, e.g., composition x_0 in Fig. 3.29), a minimal total free energy

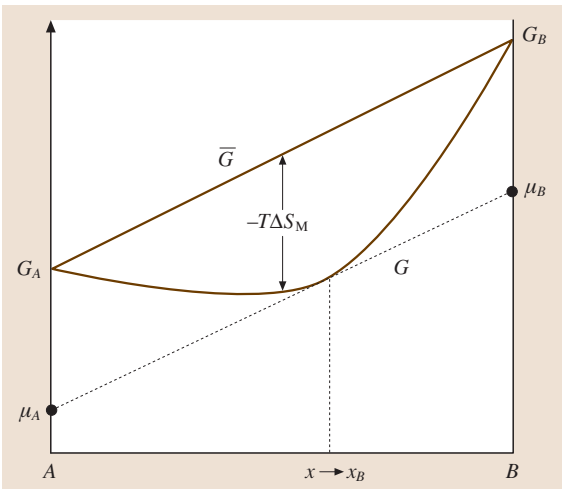


Fig. 3.28 The molar free energy of a system consisting of two components A and B before mixing ($\bar{G}$) and for an ideal solution (G)

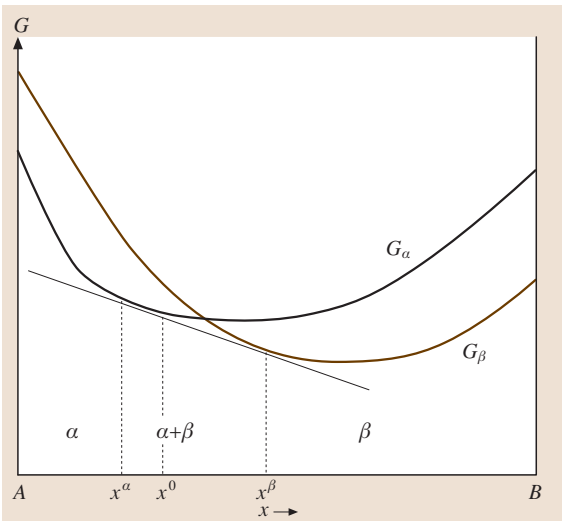


Fig. 3.29 Molar free energy curves for two phases α and β. At equilibrium, alloy x^0 has a minimum free energy when it is composed of a mixture of x^α and x^β

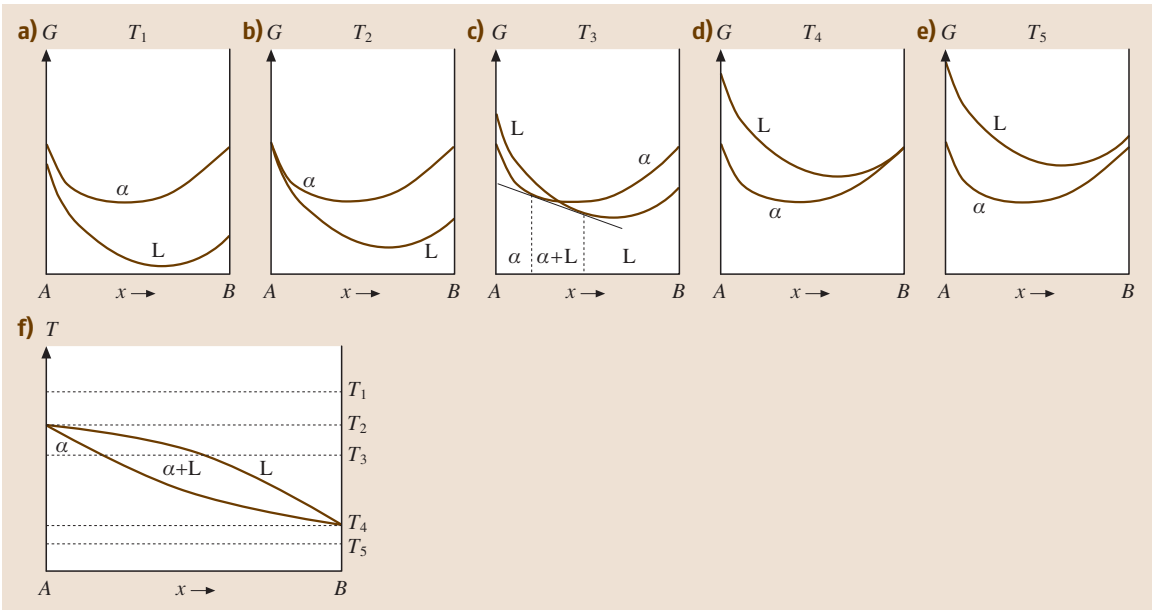


Fig. 3.30a–f The derivation of a phase diagram with complete miscibility in the liquid (L) and solid (α) state **(f)**: **(a)** at T_1 all compositions are liquid, **(b)** T_2 represents the melting point of element A, **(c)** at T_3 a two-phase region of α and L exists, **(d)** T_4 is the melting point of element β, **(e)** all compositions are solid at T_5

can be achieved by separating the atoms into the two phases α and β with equilibrium compositions x^α and x^β. From Fig. 3.29 it can be concluded that heterogeneous equilibrium between the two phases requires that the tangents to each G curve at x^α and x^β lie on a common line (the *common tangent rule*). In other words each component must have the same chemical potential in the two phases and

$$\mu_A^\alpha = \mu_A^\beta\,, \quad \mu_B^\alpha = \mu_B^\beta\,. \tag{3.32}$$

Note that the same rule applies when $G(x)$ curves are compared for liquid and solid phases.

Binary Phase Diagram with Complete Miscibility

The simplest case conceivable is when the two components A and B are completely miscible in both the liquid and solid states and both are ideal solutions. Then, the free energy curves for the liquid and solid phases vary with temperature according to Fig. 3.30. At T_1 the liquid phase is thermodynamically stable over the whole composition range (Fig. 3.30a) thus $G_\alpha > G_L$. With decreasing temperature one approaches the melting points T_2 and T_4 of pure A and B (see Fig. 3.30b,d, respectively), where the $G(x)$ curves meet in a single point on either side of the diagram. These points are plotted on the temperature axes for pure A and B, respectively, in the *equilibrium phase diagram* (Fig. 3.30f). With a further decrease in temperature (T_5) the solid phase is stable for all compositions (Fig. 3.30e), and $G_\alpha < G_L$. In the temperature interval between T_2 and T_4, the common tangent rule indicates a two-phase region with coexisting solid and liquid phase (see the two

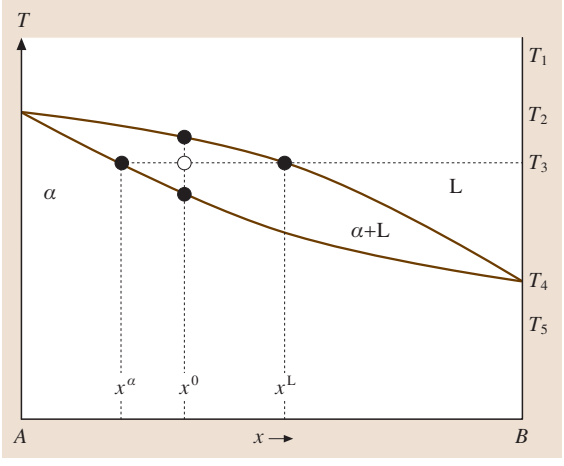


Fig. 3.31 The lever arm rule for estimating the molar fraction of solid and liquid phase in a two-phase region

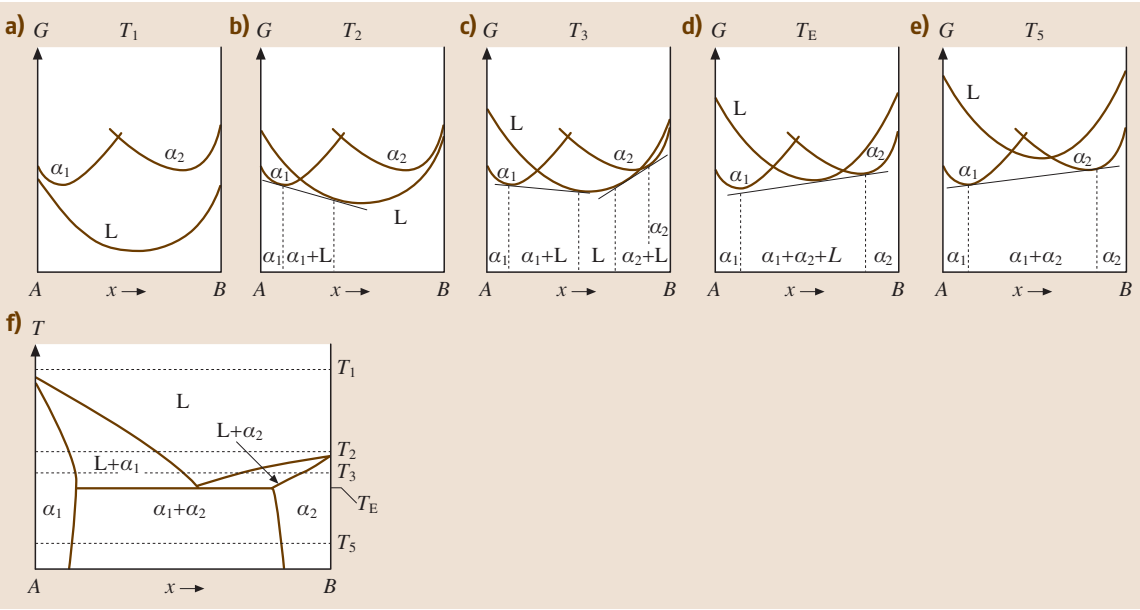


Fig. 3.32a–f The derivation of a eutectic phase diagram with a miscibility gap and limited solubility for the A-rich solid solution α_1 and the B-rich solid solution α_2 **(f)**: **(a–e)** show G–x-curves of the solid phases α_1 and α_2, respectively, and the liquid phase L at temperatures T_1–T_3, T_E and T_5. For further details see text

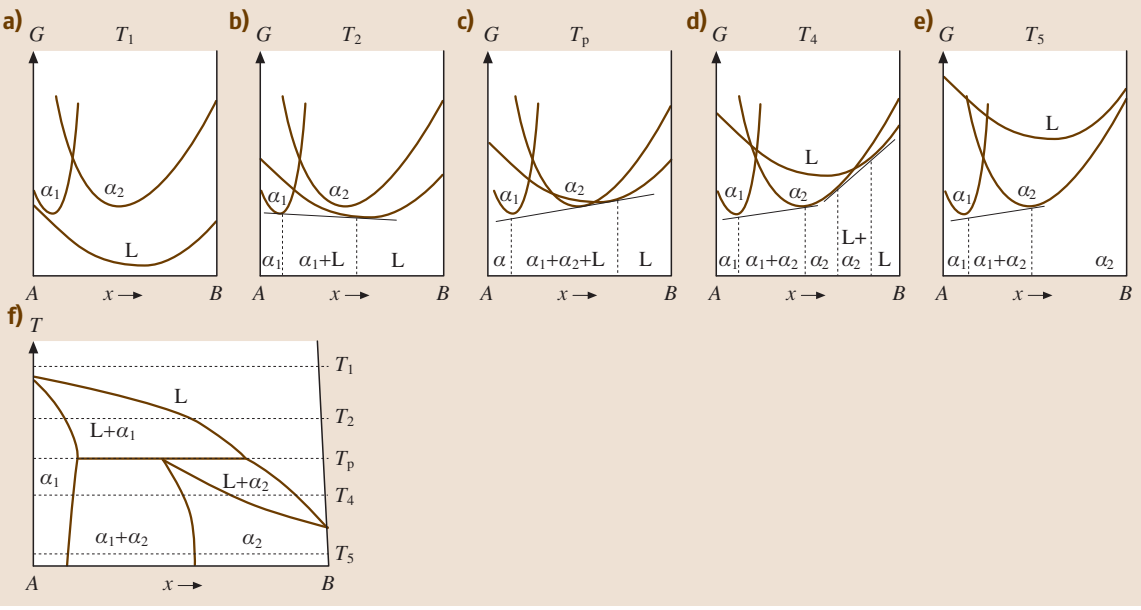


Fig. 3.33a–f The derivation of a peritectic phase diagram with a miscibility gap and limited solubility for the A-rich solid solution α_1 and the B-rich solid solution α_2 **(f)**: **(a–e)** show G–x-curves of the solid phases α_1 and α_2, respectively, and the liquid phase L at temperatures T_1, T_2, T_p T_4 and T_5. For further details see text

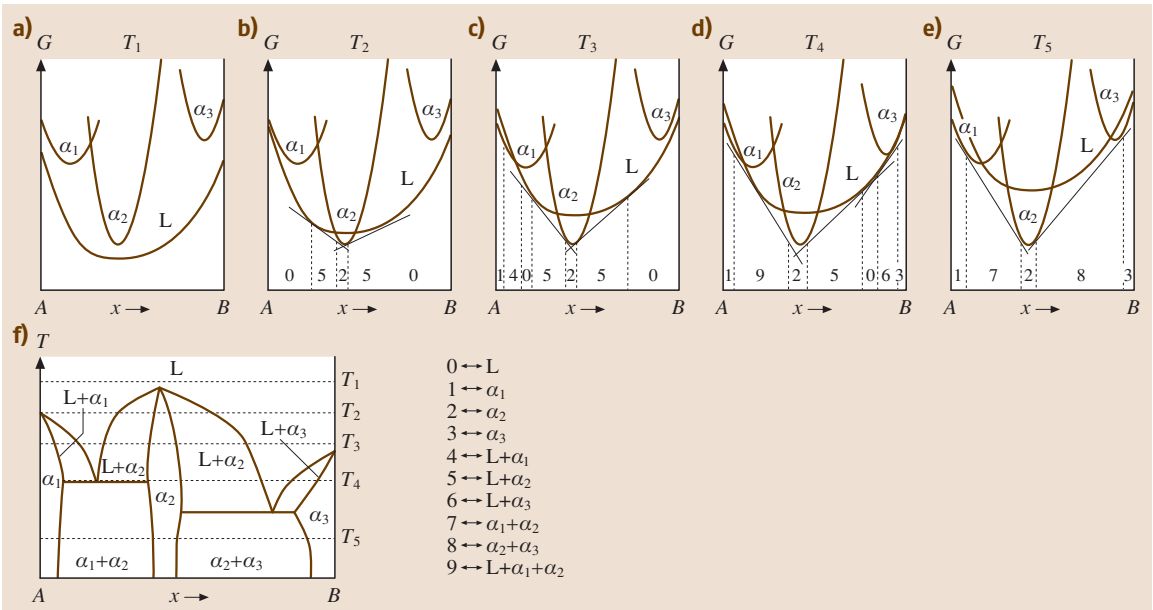


Fig. 3.34a–f The derivation of a complex system with an ordered intermetallic phase α_2 **(f)**: **(a–e)** show G–x-curves of the solid phases α_1 to α_3 and the liquid phase L at temperatures T_1 to T_5

vertical dashed lines at T_3 in Fig. 3.30c). These points are transferred into the equilibrium phase diagram at T_3.

The Lever Arm Rule

Figure 3.31 shows an enlarged view of the binary phase diagram with complete miscibility derived in Fig. 3.30. The region where the two phases coexist is limited by the two curved lines, of which the upper one separates this region from the liquid phase. Consequently, this line is called the *liquidus line*. The lower line separates the two-phase region from the solid phase and is thus called the *solidus line*. For any given temperature T_3 and overall composition x_0 within the two phase region it may now be interesting to know the amount of liquid and solid phase, respectively, as well as their concentration (or, more precisely, their molar fractions). Drawing a horizontal straight line at T_3 the intersection with liquidus and solidus line yields the concentration x^{L} and x^{α} of the liquid and the solid phase, respectively, within the two-phase region. The amount of α-phase f^{α} can be graphically determined utilizing the lever arm rule

$$f^{\alpha} = \frac{x^{\mathrm{L}} - x^{0}}{x^{\mathrm{L}} - x^{\alpha}} \,. \tag{3.33}$$

Analogously, one obtains for the amount of liquid phase f^{L}

$$f^{\mathrm{L}} = \frac{x^{0} - x^{\alpha}}{x^{\mathrm{L}} - x^{\alpha}} \,, \tag{3.34}$$

and finally for the ratio between the solid and liquid phase

$$\frac{f^{\alpha}}{f^{\mathrm{L}}} = \frac{x^{\mathrm{L}} - x^{0}}{x^{0} - x^{\alpha}} \,. \tag{3.35}$$

Equations (3.33–3.35) hold in general for all two phase regions in binary phase diagrams, e.g., for the miscibility gaps (or the two-phase regions a1 and a2 in Figs. 3.32–3.34).

Eutectic Systems

Figure 3.32 exemplifies the situation where the liquid phase is approximately ideal but the solid phase α decomposes into two solid solutions α_1 and α_2, i. e., the atoms A and B *dislike* each other. Rather, preferred A–A and B–B-type bondings can be found and ΔH_{M} in (3.31) becomes positive (regular solution). Therefore, at low temperatures T_5 the (combined) free energy curves of α_1 and α_2 in Fig. 3.32 assume a negative curvature in the middle and the solid solution is most stable as a mixture of the two phases α_1 and α_2. This region is called the *miscibility gap* and the lever arm rules (3.35–3.37) apply. The second effect of $\Delta H_{\mathrm{M}} > 0$ becomes obvious at the *eutectic temperature* T_{E} where one common tangent line can be applied to the $G(x)$ curves of all three phases: as a consequence, the *eutectic* composition has the lowest melting point within the system.

For a *eutectic reaction*

$$L \rightarrow \alpha_1 + \alpha_2 \,. \tag{3.36}$$

Peritectic Systems

Figure 3.33 illustrates how a *peritectic* phase transformation is related to the free energy curves. Again, $\Delta H_M > 0$ but, in contrast to the eutectic system, the $G(x)$ curves for the two solid solutions α_1 and α_2 are shifted to one side of composition relative to the liquid phase L. As for the eutectic system one common tangent line can be applied to the $G(x)$ curves of all three phases at T_P and the *peritectic reaction* is

$$\alpha_1 \rightarrow L + \alpha_2 \,, \tag{3.37}$$

which quite naturally explains why peritectic systems likely emerge when two components with substantially different melting points are alloyed.

Systems with Intermetallic Phases

The opposite type of effect arises when $\Delta H_M < 0$ and the atoms *like* each other within a certain composition range. In such systems (Fig. 3.34) melting will be more difficult in the α_2 phase because of its very deep $G(x)$ curve and a maximum melting point may appear. If the attraction between the unlike atoms is very strong and the α_2 phase extends as far as the liquid, it may be called an *ordered intermetallic* phase.

3.2 Microstructure Characterization

3.2.1 Basics

The primary characteristic of a material is its integral and percentual chemical composition, that is, e.g., for metals, the chemical elements, for polymer materials the types of polymers and possible reinforcements, and for ceramics the oxides, nitrides or carbides. Starting with the chemical composition, a specific microstructure [3.18] will be generated during the solidification of a melt, the mixing of polymeric components, heat treatment, the manufacturing process (rolling, milling, deep drawing, welding), or during usage (aging, corrosion).

As pointed out in detail in Sect. 3.1.2 the (usually three-dimensional) microstructure of materials can consist of several constituents, for example, grains (or crystallites) with different crystallographic orientation (which are separated from each other by grain boundaries, Fig. 3.26) or precipitates, impurities (slags, oxides, sulfides), pores, reinforcement particles, fibres, and others.

The constituents of a microstructure are visualized by material-specific preparation and imaging methods. However, for complete characterization of a microstructure (materialography, or more specifically metallography, plastography, ceramography) more methods than microscopic imaging are often necessary. For the interpretation and understanding of a microstructure the knowledge of the presence and nature of crystallinity of the constituting phases is essential. This information is obtained by X-ray diffraction, which is a nonmicroscopic integral method. The information on the local chemical composition, the local crystal structure, and characteristic geometric parameters of the constituents is investigated by microscopic methods which differ in their generated signals, optical resolution, and contrast mechanism.

3.2.2 Crystal Structure by X-ray Diffraction

The first goal in microstructure characterization is to learn which crystalline phases are present in a material. This is achieved mainly by X-ray diffraction (XRD) [3.19, 20], which gives information on the crystal structure of constituents in a microstructure. This is possible by their crystallographic parameters: type of crystal lattice, crystal symmetry, and unit cell dimension (Sect. 3.1). Moreover, information on the perfection of the crystal lattice (number of dislocations), and from this on the degree of plastic deformation, and on the external and residual stresses acting on the lattice are also obtainable.

The theory of X-ray diffraction is based on Bragg's law, which describes how electromagnetic waves of a certain wavelength λ interfere with a regular lattice. At certain angles of incidence (θ) with regard to a set of parallel crystal planes, which are therefore called reflectors, constructive interaction takes place according to

$$n\lambda = 2d_{hkl} \sin\theta \,, \tag{3.38}$$

where n is a positive integer and d_{hkl} represents the interplanar spacing between the crystal planes that cause constructive interaction; λ is the known wavelength of the incident X-ray beam.

In XRD the specimen is irradiated by a monochromatic X-ray beam, Cu-KA_α or Cr-K$_\alpha$, which is

generated by an X-ray tube and a metallic film for monochromatization. The diffracted X-ray beam is detected by, e.g., a scintillation counter at an angle of 2θ with the incident beam and the signal is stored in a computer. Both, X-ray source and detector rotate on a circle around the sample in the center. The measuring spot typically has an area of several mm^2 – in special cases of some square micrometers. Measurements are possible on bulk specimens, powders, and on films. In materials science XRD applied to polycrystalline bulk materials is also called the *powder* diffraction method. Sample preparation is relatively easily accomplished by grinding or polishing, whereby destruction of the crystal structure by severe plastic deformation must be avoided. X-ray diffraction can be treated as an integral method for measuring the crystal structure, because usually the exposed area is composed of a number of crystallites. However, single-crystal measurements are also possible. The information depth of some ten micrometer depends on the angle of incidence, the atomic number of the sample, and the energy of the X-ray.

From the X-ray diffraction diagram, which is commonly plotted as intensity of X-ray versus 2θ (Fig. 3.35), the following information can be obtained [3.20, 21]. From the angles θ of the Bragg peaks the constants of the unit cell of the crystal can be calculated by application of (3.38). From the combination of those values the type of crystal and its symmetry can be deduced. Because a set of d values and the relative intensities of their corresponding X-ray peaks are characteristic of a certain crystalline material they are used for phase identification by comparing the measured diffraction pattern to those of phases contained in databases. The most commonly used database is the powder diffraction file (PDF) maintained and distributed by the International Center for Diffraction Data (ICDD).

In case of materials consisting of multiple phases the weight ratio of the crystalline phases of the material is calculated from the relative peak intensities. The quantification without a standard sample is based on the comparison of the peak intensities. A better way to get the ratios is to use a standard substance.

The width of the peaks gives information on the perfection of the arrangement of the atoms within the crystal lattice and on the number of dislocations resulting from plastic deformation. External and residual stresses applied to a crystal lead to dilatation or compression of the atomic distances and therefore to a shift of the diffraction peaks to greater or smaller angles. Practically the strain in a sample is measured by recording the angular shift of a given reflector as a function of angle of incidence. The measured strain is then used to calculate the stress with the help of the elastic modulus [3.22].

A preferred orientation of the crystallites in a polycrystalline material with respect to the sample coordinate system is called (crystallographic) texture. The orientation distribution can be determined by X-ray diffraction (XRD)-based texture analysis [3.23]. With this technique, pole figures are measured by recording the intensity distribution of a single reflection by tilting and rotating the sample while radiating it with an

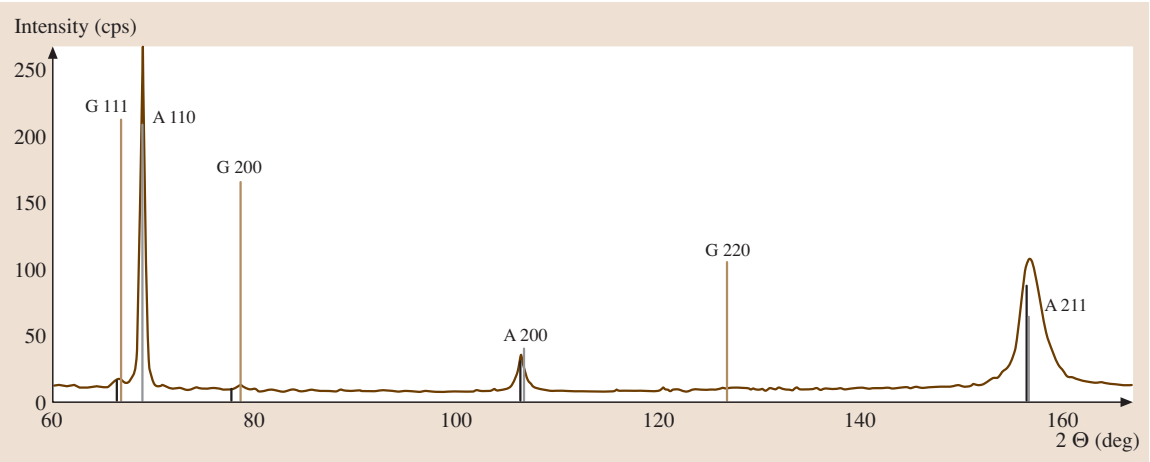


Fig. 3.35 X-ray diffraction pattern of a quenched and tempered hot-working steel 56NiCrMoV7, with metastable austenite (8%); the small peaks marked with a G represent the austenite phase whereas the large peaks marked with an A show the ferrite matrix. G and A are deduced from 8-Fe, and α-Fe, respectively

X-ray beam. In this way the orientation distribution of a single reflection, and thus for a single lattice plane, is determined [3.24].

3.2.3 Materialography

Materialography is the investigation of the microstructure of materials [3.25]. It includes specimen preparation and imaging of the microstructure, the quantification of the constituents (content, arrangement, size, shape, and orientation), as well as the local chemical and crystallographic characterization of the constituents, if necessary.

Specimen Preparation

The three-dimensional microstructure of a material is usually deduced from two-dimensional images, which are obtained by sectioning the sample. The resulting specimen is either in bulk form or thin and transparent, depending on the type of material and the goal of investigation. The whole process of specimen preparation, starting with cutting small parts from larger pieces, has to be performed without disturbing the microstructure by mechanical or thermal influences. Small specimens (wire, cross sections of sheet metal) are mounted in a resin using pans which can easily be handled and have the right size for grinding machines. Bulk samples are prepared by grinding and polishing using metallographic machines with rotating wheels. A large number of material-specific abrasives and lubricants are available [3.26]. The selection of the most suitable ones is based on the material's composition and on the mechanical properties of its constituents. Mechanical polishing is performed using a rotating wheel covered with cloth and small particle abrasives (for final polishing steps with grain size $< 1\,\mu m$), such as powders of diamond or aluminum oxide, or colloidal silicon dioxide. For further smoothing of the surface electrolytic polishing can be applied, especially for homogeneous, i. e., single phase, materials.

The prerequisite of microscopic imaging is a sufficient optical contrast, meaning that neighboring regions must show a certain difference of brightness or color. The contrast (C) is defined as the ratio of intensities I, which can be the intensity of white light (gray values) or the intensities of colors (red, green, and blue)

$$C = \frac{I_1 - I_2}{I_1}\,, \tag{3.39}$$

where $I_2 < I_1$. Contrast can already be present after polishing the samples, e.g., if black graphite is present in a bright matrix of grey cast iron, colored grains in copper alloys and mineralic materials, and contours due to different abrasion of constituents. In most cases, however, the contrast has to be developed by means of chemical or physical etching [3.27]. Chemical etchings

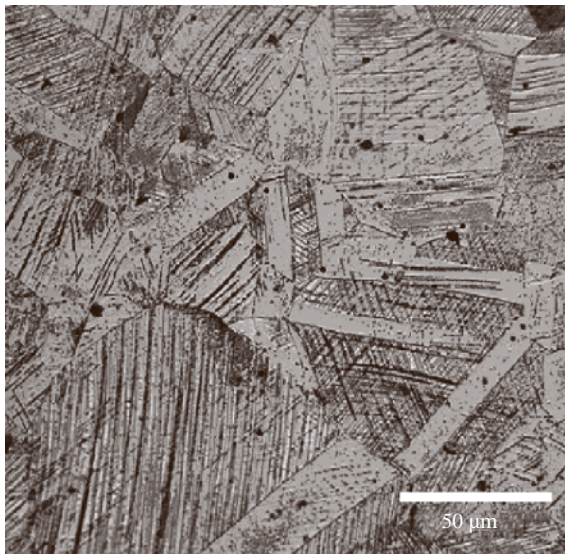


Fig. 3.36 Grain-boundary etching of an austenitic CrNi steel; the large number of twins is due to severe plastic deformation; light optical micrograph

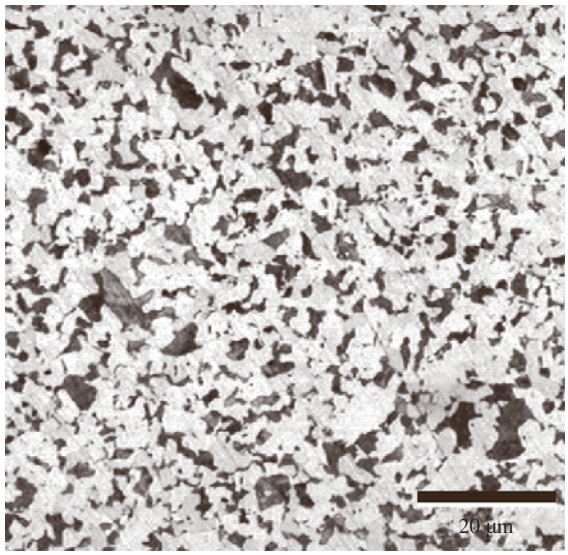


Fig. 3.37 Microstructure of a carbon steel (0.35% C), etched with 3% HNO_3; light microscopy of a polished and etched metallographic cross-section

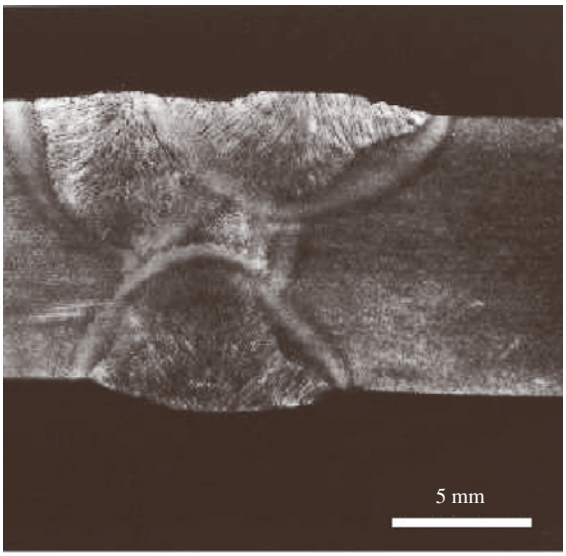


Fig. 3.38 Microstructure of a welding; macroscopy of an etched specimen

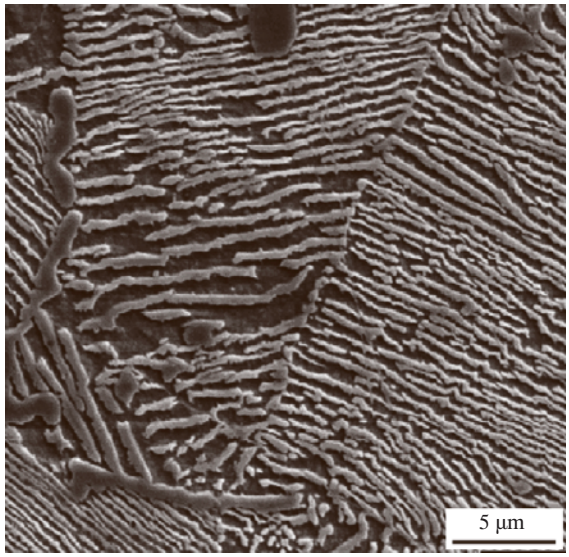


Fig. 3.39 Scanning electron microscopy (SEM) image of pearlite in plain carbon steel; secondary electron (SE) image

are water-based acidic or basic solutions or complex solutions of salts, sometimes containing organic substances. Grain boundary etching is usually applied to microstructures consisting only of one constituent (Fig. 3.36), where the etching agent reacts preferentially with the more reactive grain boundaries. Large differences in the etching rate of the constituents of a microstructure generate slopes at the grain boundaries between different constituents, which gives also a grain boundary contrast. For some etching agents the ablation depends on the crystallographic orientation of the grains and as a result topographies with different light-scattering capability are developed. If a grain consists of two phases, such as pearlite (consisting of ferrite and cementite), one of them can be selectively etched, leaving a light-scattering topography of pearlite grains, which are dark under the microscope, as compared with the brighter ferrite grains in a carbon steel (Fig. 3.37, compare also Fig. 3.39).

Physical etching methods are based on a selective ablation of constituents by a plasma generated in a glow discharge apparatus or by ion beam bombardment, for example in a focused ion beam (FIB) instrument (see later).

For light microscopy of polymer materials, transparent specimens are prepared by cutting lamellae, using a microtome with a glass or diamond knife, from the sample. The specimens are some micrometers thick and are positioned between a glass microscope slide and a cover glass by adding a drop of immersion oil to keep off air bubbles and to increase the refractive index of the interspace. Easily plastic deformable polymers, such as polyethylene, are cut at low temperatures (at −70 °C or lower) with a cryomicrotome. From polymer matrix composites thin transparent specimens are obtained by grinding and polishing small pieces which are glued to a glass strip.

Microscopy of the Microstructure

For some metallographic samples it is sufficient to image the specimen with no or only little magnification. This macrometallography is used, e.g., for the inspection of the microstructure of welds (Fig. 3.38).

In most cases, however, microscopy is necessary to visualize the microstructure. The most commonly used method is reflection light microscopy of bulk specimen. The contrast, as mentioned above, is based on the different reflection capability or color intensities of the constituents. If sufficient contrast cannot be obtained by specimen preparation, other imaging modes can be used, such as light microscopy with polarized light for aluminum and magnesium alloys, or differential interference contrast microscopy (DIC) for refractory metals (Mo, W, V).

Inverted microscopes are used for bulk metallographic samples, because they allow easy positioning of the specimen on the microscope stage with the viewed

surface exactly perpendicular to the optical axis. This is a basic requirement to have all parts of the viewed area in focus. Images are captured by a charge-coupled device (CCD) camera and a computer whereby easy-to-handle software is useful, and should allow calibration, setting scale markers (micron bar), and some interactive distance measurements. Patterns of microscope calibration standards are imaged for the calibration of the magnification of a selected microscope configuration. As a calibration value the pixel size, as micrometer per edge length of square pixels, is stored with the image. A micron bar can be placed permanently into the image if necessary, but one has to be careful if the micrograph is used for automatic image analysis afterwards.

In some cases dark-field microscopy, in which the diffuse reflected light is detected instead of the directly reflected light, gives better visibility of small objects. The lateral resolution of light microscopy is 0.25 μm at best (due to the wavelength of visible light). Best values are obtained when a substance with a large refractive index (immersion oil) is placed between the specimen and the objective.

For higher resolutions (and magnifications) than are possible with light optical methods *electron microscopy* is a method widely applied in metallography. In addition, it allows complementary information on the local chemical composition and the crystal structure to be obtained. *Scanning electron microscopy* (SEM) is used for imaging metallographically prepared surfaces of bulk samples, and *transmission electron microscopy* (TEM) is used for imaging thin foils which are transparent to electrons. In both instruments, the electrons are emitted from an electron gun, accelerated in an electric field (0.5–25 kV in SEM, and 80–400 kV – in some cases over 1 MV – in TEM) towards the anode and then formed to a small beam (with a diameter of a few nanometer) by means of an electron optical system. High vacuum is necessary all along the electron path to prevent collisions of the electrons with gas molecules.

In an SEM [3.28, 29] the specimen, mounted on a multi-axis stage in the specimen chamber, is scanned with the focused electron beam. The emitted secondary electrons (SE) and backscattered electrons (BSE) are registered by detectors which are mounted above the specimen and the signal intensities are stored as digital grey value images. The SE detector is a scintillator–photomultiplier system and for BSE a scintillator or a semiconductor detector can be used.

The best resolution is achievable with the SE signal, and can be as good as 1 nm for suitable instrument parameters and specimen constitution. The information depth is some tens of nanometers for the SE mode. For imaging of very small particles or thin layers, especially if they consist of low-atomic-number elements, the emission depth can be lowered by applying a lower accelerating voltage. With SE, a topography contrast is generated, which is based on the dependency of the SE intensity on the incident angle between the electron beam and the imaged surface area (Fig. 3.39).

With BSE a composition contrast image can be obtained, because the intensity of the BSE emission is related to the atomic number of the material. Regions containing higher-atomic-number elements are brighter than those composed of lower-atomic-number elements (Fig. 3.40). Even atomic number differences smaller then unity can result in a contrast, which is in many cases good enough for imaging the microstructure of polished, but unetched, specimens.

SEM samples have to be stable under high-vacuum conditions. This is not the case if they contain water or other liquids which can evaporate. Therefore, in some SEMs, fitted with special vacuum devices and detectors, imaging at a pressure of up to 25 mbar is possible by the injection of water into the specimen chamber; this is known as variable-pressure SEM (VPSEM) or environmental SEM (ESEM). The resulting water partial pressure prevents the evaporation of water from the specimen and an alteration of its structure. Cooling the specimen with the aid of a cooling stage to a temper-

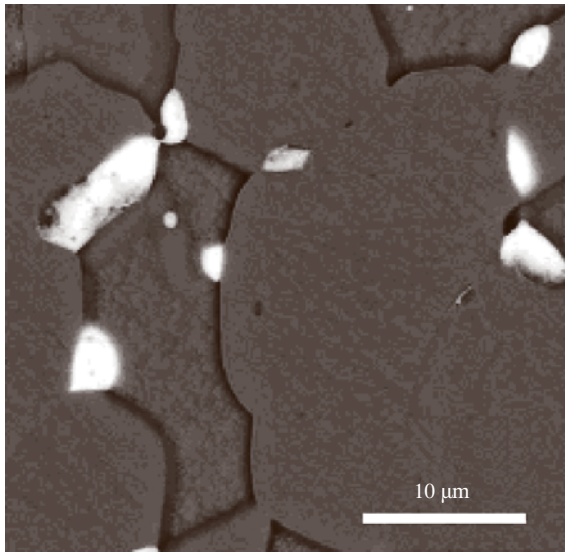


Fig. 3.40 Atomic number contrast in a SEM BSE image of brass; Pb particles are *bright* due to their higher atomic number as compared with Cu and Zn

ature just above the freezing point supports this effect. Imaging electrically nonconductive materials, such as polymers, ceramics, oxides, and mounting resins, is possible in different ways. Either they are coated with a conductive layer (Au, C, Pt, or Cr) by sputtering or evaporation, or a low accelerating voltage is applied (< 2 kV), or imaging is performed under low-vacuum conditions (at least 1 mbar), whereby ions that are generated by collision of electrons with gas atoms prevent the specimen surface from being charged.

Cross sections are commonly prepared for microstructural investigation of subsurface regions and of thin surface layers. The edge of the specimen has to be preserved to prevent its rounding and the ablation of thin layers during grinding and polishing. Often resins filled with hard particles are used, or a metal is plated on the sample surface before mounting; chemical deposition of Ni is preferred. A good alternative for the inspection of subsurface regions is cross sectioning with ion beams. Larger areas (up to some millimeters edge length) are cut with broad beams of Ar [3.30]. Target preparation of cross sections is performed using focused ion beam (FIB) instruments [3.31], in which a Ga^+ ion beam (0.5–30 kV accelerating voltage, 7 nm diameter) is scanned over the specimen. The ion bombardment results in a milling effect. Preparation is possible at any region of interest at the specimen surface by milling a stair-shaped trench, typically 20 µm wide and deep. The cross section is imaged after the specimen is tilted (Fig. 3.41). The edge of the trench is protected by a Pt strip, which is deposited before the milling by ion-induced decomposition of a metalorganic Pt compound fed into the specimen chamber through a small tube. Imaging is possible in a FIB by means of secondary ions (SI) and the ion-induced secondary electrons (iiSE), respectively. The latter give both topographical and compositional contrast. Some crystalline materials show good orientation contrast due to the channeling effect [3.32] and the microstructure is visible without etching (Fig. 3.42).

Modern instruments combine the functions of SEM and FIB. The SEM mode is used for conventional imaging with electrons, even during ion milling steps, and for charge neutralization. An energy dispersive X-ray spectrometer (EDX) and a camera for electron backscatter diffraction (EBSD) imaging (see later) can be additionally fitted to such an instrument. Thus, the real three-dimensional chemical composition, crystal structure, and microstructure of a sample can be obtained by slice-milling the wall of a cross section in small steps (50 nm to a few microns) and subsequent reconstruction of the microstructure from the resulting EDS and EBSD image series.

TEM [3.33] is used for the investigation of microstructural constituents smaller than about 50 nm in the conventional mode (CTEM) or the scanning mode (STEM), whereby a resolution of 0.1 nm can be achieved with dedicated instruments. The specimen has to be electron transparent with a thickness of 20–1000 nm, depending on the electron energy and

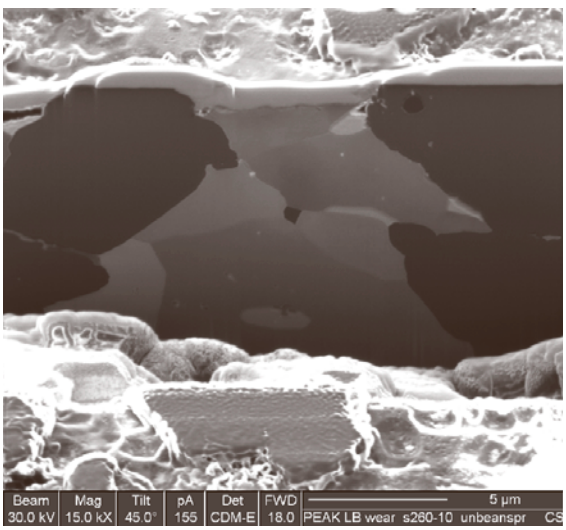


Fig. 3.41 Cross-section prepared using a focused ion beam (FIB); Al alloy, edge protected by a Pt strip, iiSE image

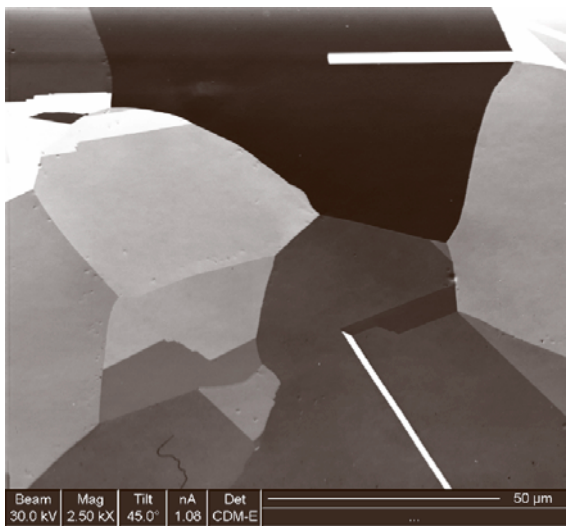


Fig. 3.42 Crystal orientation contrast due to the ion channeling effect in Cu; FIB iiSE image

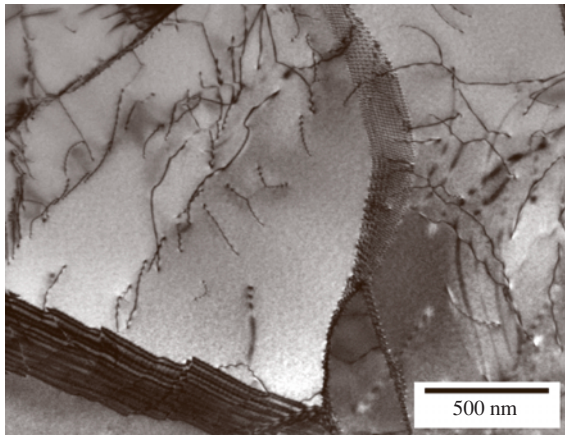


Fig. 3.43 Single dislocations within grains and dislocations forming subgrain boundaries; Mo alloy; multibeam TEM image of a 180 nm-thick specimen; 200 kV accelerating voltage

the specimen composition. Areal ion beam milling, electrolytic thinning and ultramicrotomy are common methods for specimen preparation. Usually, the specimen is mounted on a 3 mm-diameter Cu grit, which is fixed to the TEM specimen holder. Target preparation, starting with a bulk specimen, is possible by means of FIB milling. One approach is to mill a trench on either side of the region of interest, followed by cutting free the resulting lamellae and its transfer to a TEM grit with the help of a nanomanipulator.

The contrast in TEM imaging depends on different materials properties and imaging conditions. Mass–thickness contrast is based on differences in elemental composition and in thickness of the corresponding transmitted region. Diffraction contrast appears, if crystal planes are oriented in such a way that they give rise to Bragg diffraction (3.38) (Fig. 3.43). Analysis of the local chemistry of a sample in TEM is possible by means of EDX and *electron energy-loss spectroscopy* (EELS) [3.34] with a resolution of a few nanometers.

Local Chemical Analysis

Local chemical analysis is a mandatory tool to identify microstructural features such as grains, precipitates, particles, and corrosion products, or to register concentration profiles. For this purpose spectroscopy of X-rays, emitted as a result of the electron bombardment, is performed in a SEM or a TEM (*electron probe microanalysis*, EPMA) [3.35]. In most cases *energy-dispersive X-ray spectroscopy* (EDS, EDX) is used with a semiconductor detector (Si – Li or Ge) which

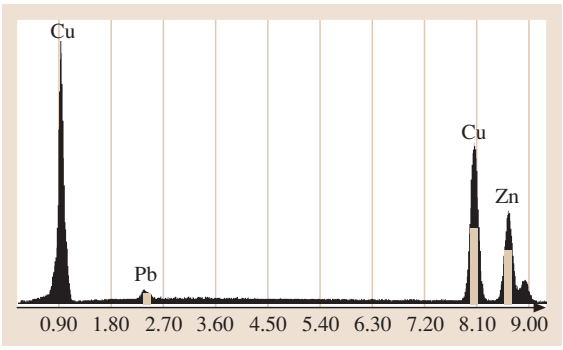


Fig. 3.44 Energy dispersive X-ray spectrum (EDX) of brass: element-specific peaks and energy windows for element mapping

is connected to a multichannel analyzer. The resulting X-ray spectrum (Fig. 3.44) gives information on the presence of chemical elements represented by the element-characteristic energy for X-ray emission. The quantitative composition is calculated from the intensities of the peaks, whereby some correcting parameters have to be taken into account [3.36]. The X-ray spectrum can represent the average elemental composition of a larger scanned area (up to 1 × 1 mm) or of a spot as small as about 0.5 µm diameter, which is the lateral resolution of EDX measurements. With a line scan the intensity of an element-specific peak (energy win-

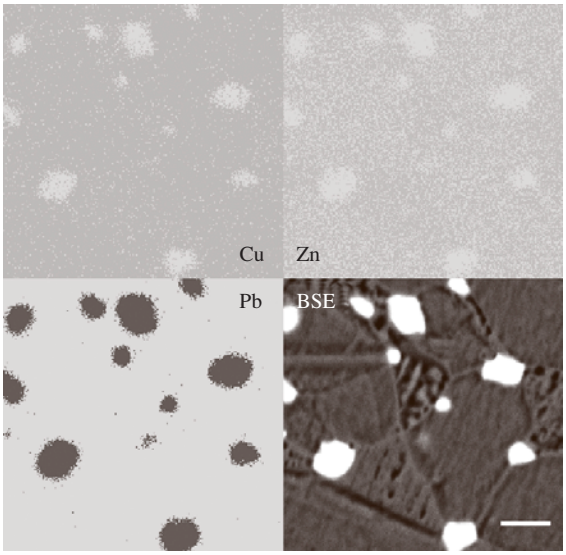


Fig. 3.45 EDX element map showing the presence of Cu, Zn, and Pb and the BSE image of the microstructure of brass

dow) is registered and from that the concentration of this chemical element can be determined along a preselected line. Extending this method to an area of interest yields so-called element mapping (Fig. 3.45).

Typically, EDX measurements in a SEM have a lateral and a depth resolution of 0.5 μm for high-atomic-number elements, and up to 10 μm for low-atomic-number elements (graphite, polymers), respectively, and relative errors of 3–8%. Better resolution can be obtained if the analysis is performed on thin specimens (≈ 100 nm thick) in both a SEM or a TEM. Elements can be analyzed qualitatively starting with the atomic number of ^{5}B whereas quantitative results can be obtained for elements starting from ^{11}Na. *Wavelength-dispersive X-ray spectroscopy* (WDS, WDX), using one or more crystal spectrometers attached to a SEM, allows the quantification of low-atomic-number elements (B, C, N, and O) and the analysis of trace elements. Because WDX cannot be used in a TEM, EELS is the alternative method of interest here.

Chemical Analysis of Thin Layers

Methods suitable for the chemical analysis of thin layers (in the nanometer thickness range), for measuring the concentration profile within such layers, and for interface layers must possess a very small information depth. Layers of interest are, e.g., sputtered or plasma-assisted coatings, corrosion layers, tribological reaction layers, and grain boundaries. Methods most used for the analysis of engineering materials are *scanning Auger electron spectroscopy* (SAM), *X-ray-exited photoelectron spectroscopy* (XPS)/ electron spectroscopy for chemical analysis (ESCA), and *secondary-ion mass spectroscopy* (SIMS) [3.35]. The lateral resolution ranges from some nanometers (SAM, SIMS) to some microns (XPS). Concentration–depth profiles are available during spectroscopy with a resolution of a few nanometers by simultaneous sputtering of the specimen with accelerated ions (O^+, Ar^+, Ga^+, etc.).

Local Measurement of the Crystal Structure

Knowing the crystal structure locally in a microstructure, for example, of a single grain or a specific precipitate is of interest for the following reasons:

1. In cases when the EDX analysis is not able to discriminate between chemically similar phases, determining the crystal structure may support phase identification.
2. Crystallographic orientation of single grains with respect to the specimen coordinates, for example, with respect to the rolling direction of sheet metal, can influence many properties significantly, such as deformation behavior, corrosion, electrical conductivity, etc.

The local crystal structure is obtained by electron diffraction with different resolutions in a TEM (< 1 nm) or SEM (> 20 nm) by applying Bragg's law (3.38). In a TEM electron diffraction of a single grain gives rise to a point pattern (Fig. 3.46) from which the relevant crystal parameters (crystal structure, symmetry, unit cell dimensions) can be deduced. It is noteworthy here that TEM has the implication that only a few grains or particles can be investigated and that sample preparation may become a difficult and tedious task.

In an SEM *electron backscatter diffraction* (EBSD) [3.37, 38] patterns are registered by a combination of a phosphor screen and a CCD camera fitted to the specimen chamber. In the pattern (Fig. 3.47) each of the so-called Kikuchi bands represents a pair of lattice planes with their width corresponding to the lattice plane spacing. From the EBSD pattern the crystal structure, symmetry, and the crystallographic orientation of a single grain can be calculated using commercial software. This method is also known as orientation imaging microscopy (OIM) [3.38]. Note, that image quality (sharpness) is deteriorated with an increasing number of dislocations within a grain, in other words with the de-

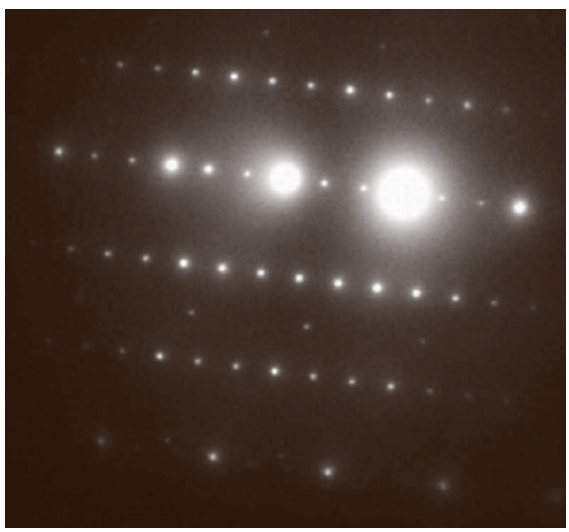


Fig. 3.46 Electron diffraction pattern of a Ni alloy obtained in a TEM at 200 kV; the small spots are superlattice peaks stemming from coherent and ordered precipitates embedded in a disordered fcc matrix

gree of plastic deformation. Since the information depth of EBSD is about 50 nm, the investigated surfaces have to be prepared very carefully by mechanical and occasionally electrochemical polishing, without disturbing the microstructure by generating dislocations due to plastic deformation [3.39, 40]. Milling the surface in a FIB or with a broad-beam ion miller is an elegant alternative preparation procedure.

For phase identification [3.41] the crystal structure obtained by EBSD and the chemical composition, which is simultaneously acquired by EDX analysis, are compared to the data of known phases contained in a database such as the ICDD database (Sect. 3.2.2).

The crystallographic orientation, defined as the orientation of the coordinate system of a crystallite with respect to the coordinate system of the sample, is calculated from the orientation of the diffraction pattern in the EBSD image. For engineering materials the coordinate system of the sample often is defined by the rolling direction (RD), the transversal direction (TD), and the normal direction (ND). To obtain an orientation map, selected areas of the specimen (up to several hundred microns edge length) are scanned with step sizes between 20 and 2000 nm, depending on grain size, and the orientation for each measuring point is calculated [3.42]. The results are usually presented as an inverse pole figure, in which the orientation is color coded (Fig. 3.48). Specifically, in Fig. 3.48, the red-colored grains have an orientation in which the crystallographic direction [001] is parallel to the normal direction of the sample, and the main axes of the cubic unit cell are parallel to RD, TD, and ND.

The different crystallographic orientation of neighboring grains can be used to generate an image of the microstructure. For this purpose the difference of the crystallographic orientation of adjacent measuring points is used; for example, a difference of less than 15° can be chosen as a criterion for discriminating between large-angle grain boundaries. The result is a colored grain map (Fig. 3.49), from which a quantitative determination of grain sizes is possible.

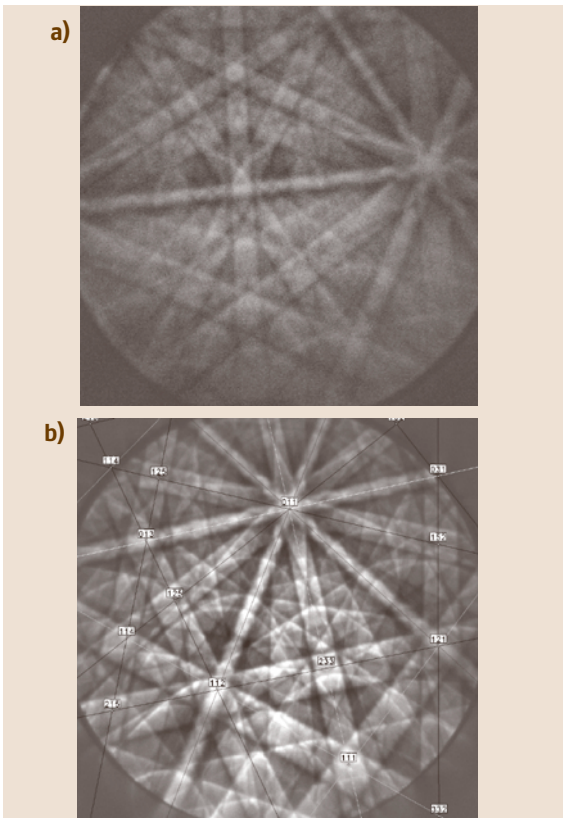


Fig. 3.47a,b Electron backscatter diffraction pattern (EBSD): of austenitic steel **(a)** and obtained from a grain in Ni with displayed zone axis directions **(b)**

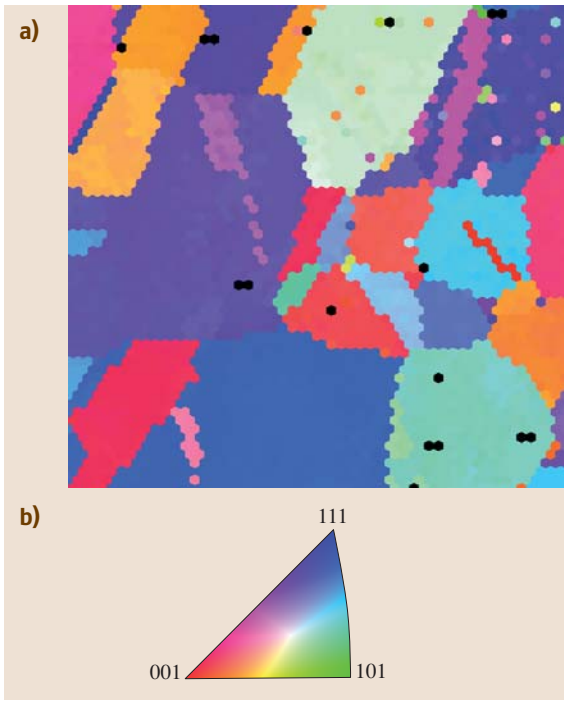


Fig. 3.48a,b Crystallographic orientation of grains in a polished section of pure Cu; inverse pole figure map **(a)** and corresponding legend **(b)**

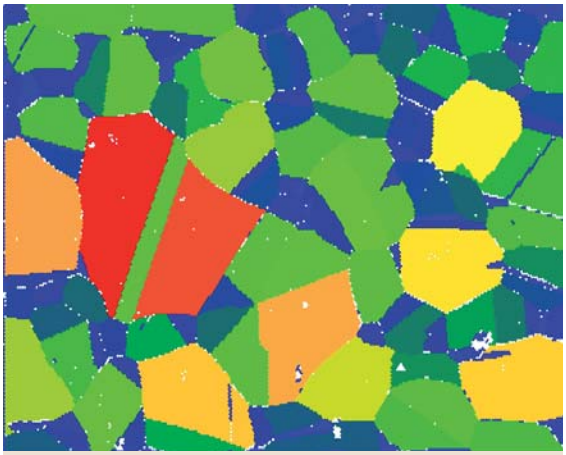


Fig. 3.49 Microstructure of an austenitic Cr–Ni steel based on EBSD measurement; different grains were defined as areas with misorientations larger than 15°

Quantification of Microstructure/Quantitative Stereology

In many cases, a quantitative analysis of the microstructure is desired, e.g., to detect small differences between microstructures for quality control purposes, or to obtain numbers for modeling of material behavior. One example is the correlation of the grain size with the strength of a material as described by the Hall–Petch equation (3.10).

Microstructure quantification can be performed by image comparison using standard charts, by manual measurement or counting, and with the help of digital image analysis software. Standards for microstructure quantification describe the specimen preparation as well as the measuring procedure, the necessary equipment, and the form of test report.

The study of the three-dimensional microstructure using images of two-dimensional sections through the structure is known as stereology [3.43, 44]. A basic stereological parameter [3.45,46] is the volume fraction $V_V(\alpha)$ of a constituent (α), expressed as the ratio of the constituent volume $V(\alpha)$ and the testing volume V_t

$$V_V(\alpha) = \frac{V(\alpha)}{V_t} \,. \tag{3.40}$$

The volume fraction equals that fraction which can be obtained from the corresponding equilibrium phase diagram, where the volume fraction has to be calculated from the weight fraction using the densities of the constituents. The volume fraction can be estimated using the fundamental stereological equation

$$V_V(\alpha) = A_A(\alpha) = L_L(\alpha) = P_P(\alpha) \tag{3.41}$$

from the areal fraction $A_A(\alpha)$ to be determined by digital image analysis, from the line fraction $L_L(\alpha)$ by digital image analysis or by measuring the sum of the length of all segments $L(\alpha)$ of test lines L_t which lie within the grains of constituent (α), or from the point fraction $P_P(\alpha)$. The latter is estimated by a manual point-counting procedure, in which a point grid is placed on the micrograph and the total number of points in the testing area P_t and the number of points hitting the constituent of interest $P(\alpha)$ are counted. This method has been standardized for steel [3.47] and duplex steel [3.48]. Rules for the use of automatic image analysis to determine the volume fraction of constituents are also described [3.49].

Another important stereological parameter is the surface density as an equivalent to the grain size. It is calculated as the sum of the surfaces (boundaries) of all grains $S(\alpha)$ of a constituent in a given test volume V_t

$$S_V(\alpha) = S(\alpha)/V_t \quad \left(\mathrm{m}^{-1}\right) . \tag{3.42}$$

$S_V(\alpha)$ is obtained by counting the number of intersections of test lines of total length L_t with grain boundaries P

$$S_V(\alpha) = \frac{2P}{L_t} = 2P_L \quad \left(\mathrm{m}^{-1}\right) . \tag{3.43}$$

Several parameters are known for the quantification of grains, for example, size, shape, orientation, and arrangement. A simple procedure for determining the average grain size is accomplished by comparing micrographs of the sample at a given magnification to a standard chart of grains, as standardized for graphite in grey cast iron [3.50] and for copper and its alloys [3.51], or by measuring the mean intercept length $\bar{L}_S$ (intercept method). The latter has been standardized for ferritic and austenitic steel [3.52, 53]. It is estimated by laying test lines of total length L_t over the micrograph and measuring the mean length of the segments $\bar{L}_S$ within the grains, or the number of intersections of test lines with grain boundaries P in order to calculate $\bar{L}_S$ as

$$\bar{L}_S = \frac{L_t}{P} \,. \tag{3.44}$$

The surface density is related to the mean intercept length by

$$S_V = \frac{2}{\bar{L}_S} = 2P_L \,. \tag{3.45}$$

A relatively simple and quick procedure to describe the grain size, particularly applied for ferritic and austenitic steel, is to use the grain size number G [3.53]. This is obtained by image comparison or by estimating the number of grains m for a unit area of 1 mm^2 of the specimen. In this procedure a circle is drawn into the micrograph and the number of grains within the circle n_1 and those touching the perimeter n_2 are counted. With the microscopic magnification g and the circle radius r the number of grains per mm^2 of the specimen surface is calculated as

$$m = \frac{\left(n_1 + \frac{n_2}{2}\right) g^2}{r^2 \pi} \,. \tag{3.46}$$

The grain size index G is calculated from [3.52]

$$m = 8 \times 2^G \,, \tag{3.47}$$

and is related to the mean intercept length $\bar{L}_S$ and the surface density S_V through

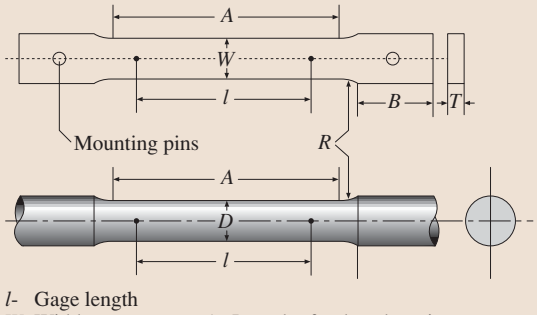
$$\begin{aligned} G &= 16.64 - 6.64 \log\left(\bar{L}_S\right) \\ &= 16.64 - 6.64 \log\left(\frac{2}{S_V}\right) \\ &= 16.64 - 6.64 \log\left(P_L\right) . \end{aligned} \tag{3.48}$$

The shape of single grains and particles is estimated by image comparison or by digital image analysis. With the latter method, geometric parameters for each particle such as area A, perimeter P, and greatest and smallest extent d_{max} and d_{min} are measured. Shape factors are calculated by expressing the deviation from circular shape (circular form factor $f_c = 4\pi A/P^2$) and the aspect ratio $f_a = d_{max}/d_{min}$. Those form factors are needed to verify the results of heat treatment, where grains are rounded, or rolling and deep drawing processes, where grains may be elongated.

3.3 Mechanical Properties

3.3.1 Framework

Concepts such as elastic properties, fracture toughness, fatigue, plastic flow, creep, etc. all belong to the framework of *mechanical properties*. Engineers and scientists working in fields related to engineering materials require a fundamental understanding of mechanical properties. Engineers are primarily concerned with the *strength* of the material, a measure of the external force required to overcome internal forces of attraction between the fundamental building blocks of the material.



l- Gage length
W- Width
T- Thickness
R- Radius of filler
A- Length of reduced section
B- Length of grip section
D- Diameter of reduced section

Fig. 3.50 Schematics of flat and cylindrical tensile test samples with critical sample dimensions (after [3.49])

In most engineering applications, only very small deformation in a component under a given loading condition is tolerable, and strength often governs the choice of an acceptable material. For a production engineer, though, the ease of inducing permanent deformation at the expense of as little energy as possible (i. e., *malleability* and *ductility*) is the critical mechanical property for the material under consideration. Given the importance of mechanical properties, it is essential to have a range of tests to *quantify* these mechanical properties. Additionally, standardized and inexpensive tests are needed for *quality assurance*. Scientists and alloy designers routinely use mechanical tests to assess the *performance* of a new material as compared with available materials.

3.3.2 Quasistatic Mechanical Properties

Tensile Testing

In this section, we address the response of a material to the application of an external applied static (or quasistatic) force. In its simplest form, the basic description of a material is obtained by a *tension* (or pull) test. Standard procedures for sample preparation and conducting the test are described in ASTM standard E 8M-98. Accordingly, the test specimen may be plate, sheet, round, wires or pipes (Fig. 3.50) and must conform to certain guidelines in terms of sample dimensions. Different gripping mechanisms, such as wedge

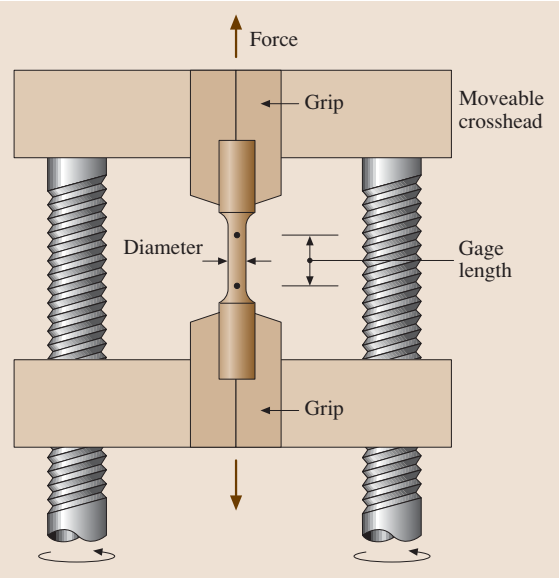


Fig. 3.51 Schematic representation of the components of a tensile testing machine (after [3.54])

grips, threads, pins or shoulders, may be considered during specimen design. It is important to emphasize that, during specimen preparation, special care must be directed towards ensuring that the reduced section of the sample is free of defects, both microstructural and machining defects, and that the specimen is representative of the bulk material.

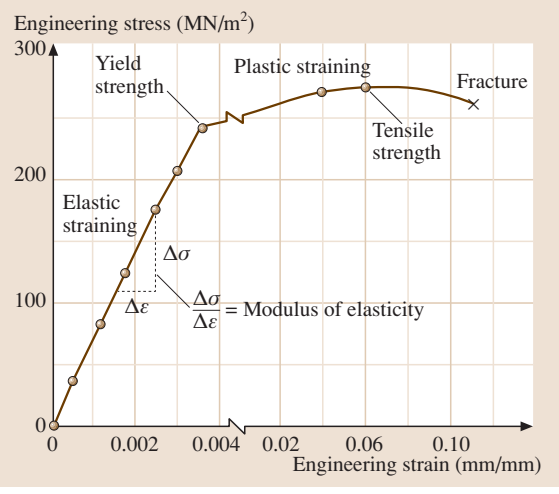


Fig. 3.52 A typical engineering stress–strain curve for a ductile material showing key mechanical properties (after [3.54])

A typical tensile testing machine (Fig. 3.51) comprises a stiff frame, a specimen gripping device, a force measuring device (or load cells), an elongation measuring device (extensometer), and a data-recording device (X–Y plotter or computer). After careful measurement of the relevant specimen dimensions, a tensile test may be run at a constant rate. The load required to produce a given elongation is recorded as the specimen is pulled and is plotted on a load–elongation chart. In order to obtain a more fundamental description of material properties, it is essential to normalize the load–elongation data for specimen geometry. To achieve this, load and elongation are converted into engineering stress and engineering strain, respectively.

Engineering stress (σ) is defined as

$$\sigma = \frac{P}{A_0}\,, \tag{3.49}$$

where P is the applied load in Newtons and A_0 is the original area of cross-section of the test specimen in square millimeters. Engineering strain (ε) is defined as

$$\varepsilon = \frac{l - l_0}{l_0} = \frac{\Delta l}{l_0}\,, \tag{3.50}$$

where l and l_0 are gage length (see Fig. 3.50 for definition) under load and original gage length, respectively.

Figure 3.52 shows a schematic engineering stress–strain diagram obtained from a tension test. The diagram is divided into two distinct regions:

1. Exclusively elastic deformation, i. e., linear and fully recoverable upon removal of load
2. (Elastic and) plastic deformation, where the latter is the nonlinear and nonrecoverable portion of total deformation

From this curve, certain key material properties can be evaluated as described below:

- *Young's modulus* E: The ratio of axial stress to corresponding strain in the elastic region. In some materials (typically polymers) the elastic region of the curve is not perfectly linear and a chord method is applied to estimate elastic modulus (Fig. 3.53).
- *Yield strength* σ_y: The stress at which it is considered that plastic elongation of the material has commenced. This stress may be specified according to one of the definitions:
  - A specified deviation from a linear stress-strain relationship, i. e., proof stress
  - A specified total extension attained or
  - Maximum and minimum engineering stresses measured during discontinuous yielding, i. e.,

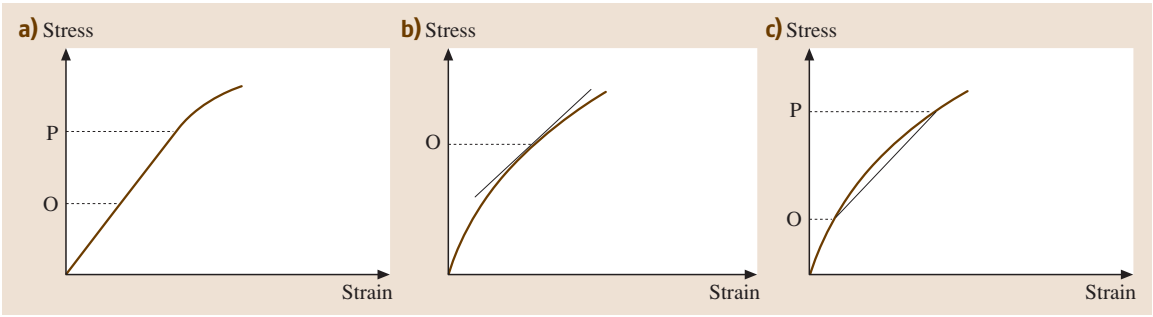


Fig. 3.53a–c Different methods to calculate Young's modulus: **(a)** from the slope of the curve between O and P below the proportional limit, **(b)** from the tangent at a given stress O, and **(c)** from the slope of the chord between stress O and P (after [3.53])

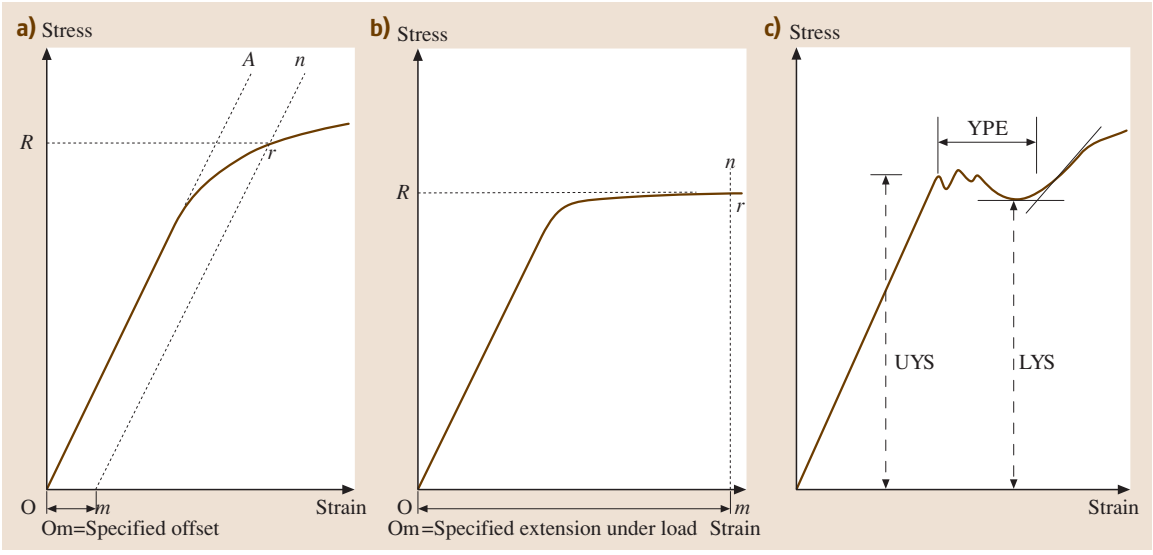


Fig. 3.54a–c Calculation of the yield stress according to **(a)** prespecified plastic offset **(b)** prespecified total strain, and **(c)** by upper and lower yield point (upper yield stress (UYS), lower yield stress (LYS), yield point elongation (YPE)) (after [3.53])

upper and lower yield point, respectively (Fig. 3.54)

The yield stress of a material may be engineered by altering grain size, and adding alloying elements and/or second phases.

- *Ultimate tensile strength (UTS)*: the maximum stress recorded during the tensile testing. After this stress level is reached, the specimen starts to show localized deformation called *necking*. Beyond this point, the engineering stress is seen to fall due to the fact that the engineering stress is defined according to original specimen dimensions. However, the true stress ($\sigma = P/A$, where A is the actual area of cross-section) continues to rise until fracture.
- *Ductility/elongation*: the ability of a material to deform before fracture under tensile load. Ductility, using this test method, is frequently quantified as percentage elongation at failure, i. e., $\varepsilon_{\text{fracture}}(\times 100)$, where $\varepsilon_{\text{fracture}}$ is the engineering strain at point of fracture.
- *Resilience and toughness*: The ability of a material to absorb energy when deformed elastically/plastically. It is defined as the area under the stress-strain curve in the elastic and plastic region, respectively.

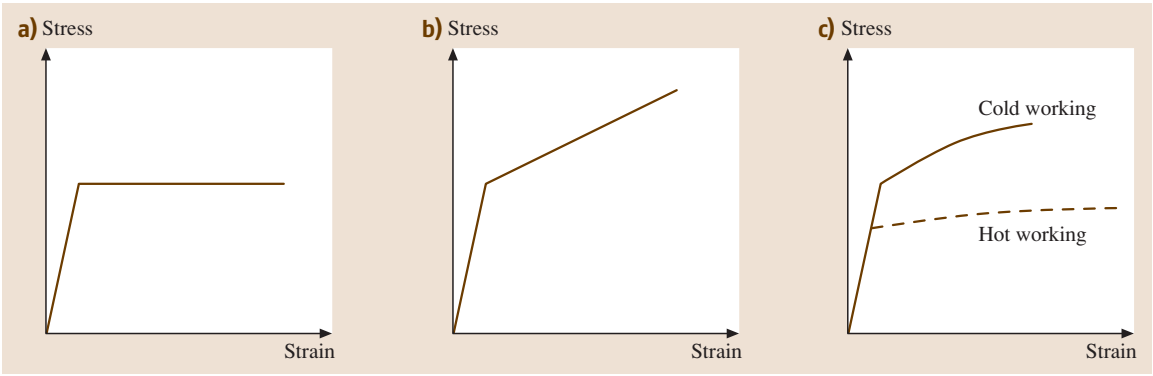


Fig. 3.55a–c Schematic representation of strain hardening in ductile metals. **(a)** elastic-ideal plastic, **(b)** elastic-plastic, and **(c)** flow curve during cold working showing strain hardening and hot-working without significant strain hardening (after [3.13])

To differentiate between elastic and plastic regions of the stress–strain curve, it is appropriate to look at the origin of strain. During elastic deformation, it is the stretching of interatomic bonds that leads to observed macroscopic strain and is linear due to the nature of interatomic forces. On the other hand, the fundamental mechanism of plastic deformation is distortion and reformation of atomic bonds. During this process the total volume of the material, however, is conserved. During plastic deformation, *dislocations* within the material become operative and *slip* due to shear stresses acting on them. For an ideal plastic, the stress required for dislocations to continue slipping is a material constant and does not depend on prior strain (Fig. 3.55a). However, in real materials, as deformation proceeds, more dislocations are generated within the material and additional driving force/stress is required for slip to proceed. This phenomenon is called *strain hardening* and is beneficially exploited during *cold working* to raise yield strength of the resultant material. Strain hardening may be overcome by *hot working* since dislocations start to become annihilated at higher temperatures ($T > 0.5T_m$, where the temperatures are calculated in Kelvin, Fig. 3.55). Furthermore, mechanical properties such as elastic modulus and tensile strength are strongly temperature dependent and decrease with increasing temperature. Ductility though is generally found to increase with increasing temperature.

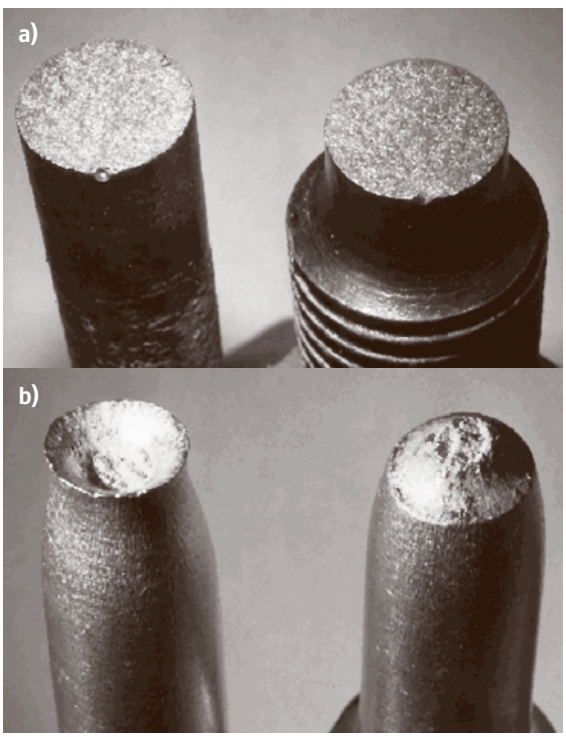


Fig. 3.56a,b Appearance of fracture surfaces after tensile testing. **(a)** Brittle fracture leads to a relatively flat surface whereas **(b)** ductile fracture shows considerable deformation prior to fracture, leading to a classical cup-and-cone arrangement

The plastic region of the true stress–true strain curve is also referred to as the *flow curve* as it is the locus of stress required to cause the metal to *flow* plastically to any given strain. The most common expression to describe the flow curve empirically takes the following form:

$$\sigma = K\varepsilon^n \,, \tag{3.51}$$

where K is the stress at $\varepsilon = 1$ and n, the strain hardening exponent, is the slope of a log–log plot of the flow curve. It can be shown mathematically that $n = \varepsilon_{\text{fracture}}$ and therefore a measure of the ductility of material.

The appearance of the surface of the specimen after fracture provides clues to the mode of fracture. A brittle fracture is accompanied by a flat (and *grainy*) surface as shown in Fig. 3.56a while ductile fracture after considerable plastic deformation has a *cup-and-cone* fracture surface (Fig. 3.56b).

Compression Testing

Mechanical properties such as yield strength, yield point, elastic modulus, and stress–strain curve may also be determined from compressive tests. This test procedure offers the possibility to test brittle and nonductile metals that fracture at low strains and avoids the complications arising out of necking. On the other hand, for certain metallic materials, *buckling* and *barreling* complicate testing (Fig. 3.57) and can be minimized by designing the samples as per specifications and using proper lubricants. Solid round/rectangular cylindrical samples (aspect ratio 0.8–10) may be used. Surface flatness and parallelism are important considerations during sample machining. After marking the gauge length and measuring the specimen dimensions, the specimen is placed in the test fixture and should be aligned carefully to ensure concentric loading. The specimen is then subjected to an increasing axial compressive load; both load and strain may be monitored either continuously or in finite increments. Relevant mechanical properties may be determined as described in Sect. 3.1.1. Compression testing is usually easier to conduct than tension test and is used more commonly at elevated temperature in plasticity or formability studies since it simulates compressive stress as is expected under rolling, forging or extrusion operation.

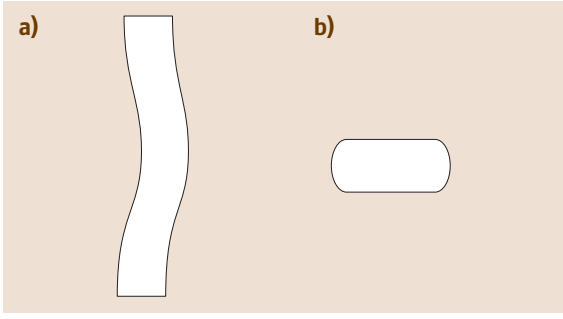


Fig. 3.57a,b Schematic representation of **(a)** buckling and **(b)** barreling during compressive testing (after [3.53])

Hardness

In general, the hardness of a material refers to its resistance to plastic deformation and is a loosely defined term. However, it is an easily measurable quantity and frequently employed in quality assurance and inspection. Standard hardness test procedure involves slowly applying an indentation to the surface of the material and measuring the relevant dimensions of the depression. Depending on the shape of the indenter and method of calculation, the following hardness tests are commonly employed.

Brinell hardness: An indenter of hardened steel or tungsten carbide *ball* with diameter D (1–10 mm) is forced into the surface of a test piece and the diameter of the indentation, d, left in the surface after removal of the test force F (100–3000 kgf) is measured. Brinell hardness (BHS or BHW) is then obtained by dividing the test force by the curved surface area of the indentation as

$$\text{BHS or HBW} = \frac{2F}{\pi D(D - \sqrt{D^2 - d^2})} . \qquad (3.52)$$

Later Meyer suggested a more rational definition of hardness based on projected area but it did not gain acceptance, despite its more fundamental nature. Meyer hardness is given as $4F/\pi d^2$ kgf/mm^2.

The *Vickers hardness* test uses a *square-based diamond pyramid* as the indenter with the included angle between the opposite faces being 136°. Due to the shape of the indenter, the Vickers hardness number (VHN or VPH) is also frequently referred to as the diamond-pyramid hardness number (DPH) and is defined as the load divided by the surface areas of indentation according to the following equation:

$$\text{DPH} = \frac{2F\sin(\theta/2)}{L^2} = \frac{1.854\,\text{F}}{L^2} , \qquad (3.53)$$

where L is average length of diameters in mm and θ is the angle between opposite faces of the diamond (= 136°). The advantage of the Vickers hardness is that it provides a continuous scale of hardness, from very soft metals to very hard materials. On the other hand, VHN is fairly sensitive to surface finish and human error.

Rockwell hardness is the most widely used hardness test in the industry due to its speed, freedom from personal error, and ability to distinguish small hardness differences in hardened steels. This test utilizes the depth of indentation, under constant load, as a measure of hardness. A minor load of 10 kg is first applied to seat the specimen, followed by the major load for the required dwell time. The depth of indentation is

automatically recorded electronically or by a dial indicator in terms of an arbitrary scale without units. THe Rockwell hardness indenter is either a 120° diamond *spheroconical* or steel *balls* 1/16–1/2 inch in diameter. Major loads of 60, 100, and 150 kg are used. Different combinations of load and indenter are used for material with different hardnesses and it is necessary to specify the combination employed when reporting Rockwell hardness. This is done by prefixing the hardness number with a letter indicating the particular combination. Hardened stress is tested on the C scale with the diamond indenter and a 150 kg major load.

Hardness testing is a very useful and reproducible method to measure and compare the mechanical strength of a material provided that sufficient precautions are taken during testing. Hardness tests are carried out on the surface of the specimen and therefore it is very important that the surface is flat, free of defects, and representative of the bulk material. Additionally there are empirical correlations available to estimate tensile strength from hardness value and to convert a result of one type of hardness test into those of a different type. However, it is important to verify these correlations for the specific class of material under consideration, though some standard conversion tables for commercial carbon and alloy steels and aluminum alloys are available. Micro- and nanohardness testing procedures are available for measuring hardness over smaller areas, while hot hardness testers are used to measure hardness at elevated temperatures.

Bend or Flexure Testing

Bend tests are used primarily for obtaining values of proof stress and modulus of elasticity in bending (E_b) as well as the ductility of relatively flexible materials such as polymers and their composites. Bend testing also provides a convenient method for characterizing the strength of the miniature components and specimens that are typical of those found in microelectronics applications.

There are two test types (Fig. 3.58): three-point flex and four-point flex. In a three-point test the area of uniform stress is quite small and concentrated under the center loading point. In a four-point test, the area of uniform stress exists between the inner span loading points (typically half the outer span length).

A flexure test produces tensile stress in the convex side of the specimen and compression stress in the concave side. This creates an area of shear stress along the midline. To ensure that the primary failure comes from tensile or compression stress the shear stress must

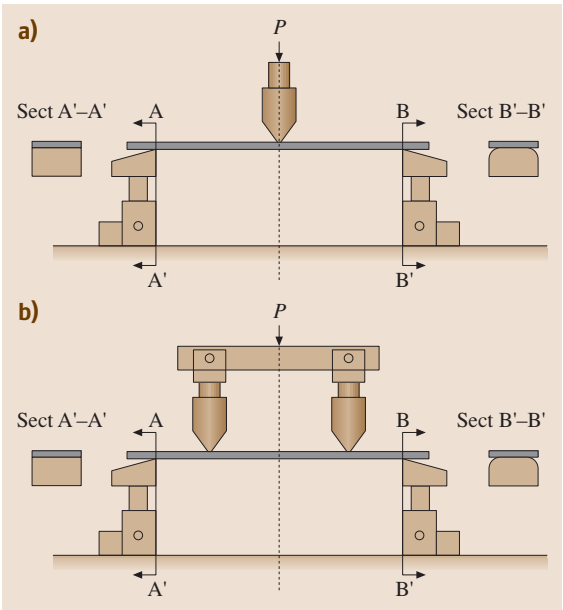


Fig. 3.58a,b Schematic representation of **(a)** three- and **(b)** four-point bend tests (after [3.53])

be minimized. This is done by controlling the span (S) to depth (d) ratio, the length of the outer span divided by the height (depth) of the specimen. For most materials $S/d = 16$ is acceptable. Some materials require $S/d = 32–64$ to keep the shear stress low enough. Usually, a rectangular cross-section of the specimen is used. E_b varies as the third power of beam thickness, and therefore uniformity in thickness is of paramount importance.

The test apparatus consists of two adjustable supports and means of measuring deflection and for applying load. The supports are generally knife-edge or convex. The load applicator is a rounded knife-edge with an included angle of 60°, applied either at mid span (for three-point testing) or symmetrically placed from the supports (for four-point testing). Elastic deflection δ is measured at the mid-span as shown in Fig. 3.58. Stress and E_b are related to applied load and deflection as follows:

$$\sigma_p = \frac{3PL}{2bh^2}\,; \quad E_b = \frac{PL^3}{4bh^3\delta}$$

for three-point bend testing (3.54)

and

$$\sigma_p = \frac{3Pa}{bh^2}\,; \quad E_b = \frac{Pa\left(3L^2 - 4a^2\right)}{4bh^3\delta}$$

for four-point bend testing (3.55)

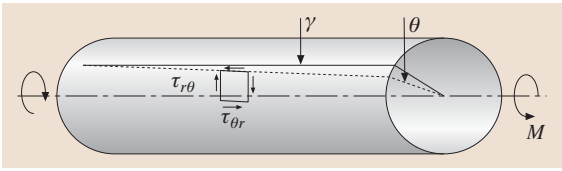


Fig. 3.59 Idealized schematic of a tubular pipe under torsion showing the shear stress and strain (after [3.13])

Torsion Testing

Torsion tests have not met with the wide acceptance and use that have been given to tensile testing. However, in many engineering applications and theoretical studies, they are of considerable importance. Torsion tests are used to determine such properties as modulus of elasticity in shear (G), torsional yield strength, and modulus of rupture. During a torsion test, measurements are made of the twisting moment, M_T and the angle of twist θ. From linear elastic mechanics (Fig. 3.59), shear stress τ and shear strain γ can be calculated according to the following relation:

$$\tau_{max} = G\gamma = \frac{M_T r}{J}\,, \quad \gamma = \frac{r\theta}{L}\,, \tag{3.56}$$

where r and L are the radius and length of the test specimen, respectively, and J is the polar moment of inertia of the area with respect to the axis of specimen. For a solid cylindrical specimen, $J = \pi r^4/2$. Because of the stress gradient across the diameter of a solid bar, it is preferable to use a tubular specimen for the determination of the shearing yield strength and elastic modulus. However, care must to be taken to avoid buckling in this case.

Beyond the torsional yield strength, shear stress is no longer a linear function of distance from the axis and the analysis becomes slightly more complicated; in this case the maximum shear at the surface is given as

$$\tau_{max} = \frac{1}{2\pi r^3}\left(\theta\frac{dM_T}{d\theta} + 3M_T\right)\,. \tag{3.57}$$

It is possible to convert the shear stress–strain curve into a tensile stress–strain curve using the following relations:

$$\sigma = \sqrt{3}\tau_{max}\,, \quad \varepsilon = \frac{\gamma}{\sqrt{3}}\,. \tag{3.58}$$

A major advantage of torsion tests over tensile tests is that fracture is delayed and it is possible to extend the flow curve to larger strains. This is of significance in the study of the plastic flow behavior of ductile materials. Torsion tests are regarded as complicated due to the considerable labor involved in converting torque–twist data to stress–strain data. However, computational resources available today considerably ease this restraint. Torsion testing provides a more fundamental description of the plasticity of metals and avoids complications such as necking and barreling associated with tension and compression tests.

As mentioned earlier, specimen design and fabrication is rather important for obtaining reliable mechanical properties from torsion tests. Specimens in the form of solid cylinders should be straight and of uniform diameter with a length equal to the gauge length plus two to four diameters. In the case of tubes, the total specimen length should be the gauge length plus at least four outside diameters. The prescribed ratio of gauge length to diameter is at least four to ten. For tubular samples, the ratio of outside diameter to wall thickness should lie between eight and ten. During testing, the twist angle is generally applied by mechanical, optical or electrical means using rings fastened to the sample. A torsiometer, fastened to the sample and the base of the machine, is used to measure the angle of twist in radians in both elastic and plastic regions.

Creep Testing

A metal subjected to constant load at elevated temperature ($> 0.5T_m$, where T_m is the absolute melting temperature) undergoes time-dependent (anelastic) deformation called creep. At these temperatures the mobility of atoms increases significantly according to Fick's laws, (3.16) and (3.18), and according to (3.17) diffusion-controlled processes have a significant effect on mechanical properties. Rate of dislocation climb, concentration and mobility concentration of vacancies, new slip systems, and grain boundary sliding are all temperature and diffusion controlled and affect the mechanical behavior of materials at high temperatures. In addition, corrosion or oxidation mechanisms, which are diffusion-rate dependent, will have an effect on the lifetime of materials at high temperatures.

Conceptually a creep test is rather simple: a force is applied to a test specimen exposed to a relatively high temperature and the dimensional change over time is measured. If a creep test is carried to its conclusion (that is, fracture of the test specimen), often without precise measurement of its dimensional change, then it is called a stress rupture test. Although conceptually quite simple, creep tests in practice are more complicated. Temperature control is critical (fluctuation must be kept to $< 0.1–0.5\,°C$). Resolution and stability of the extensometer are important concerns (for low-creep materials, displacement resolution must be on the or-

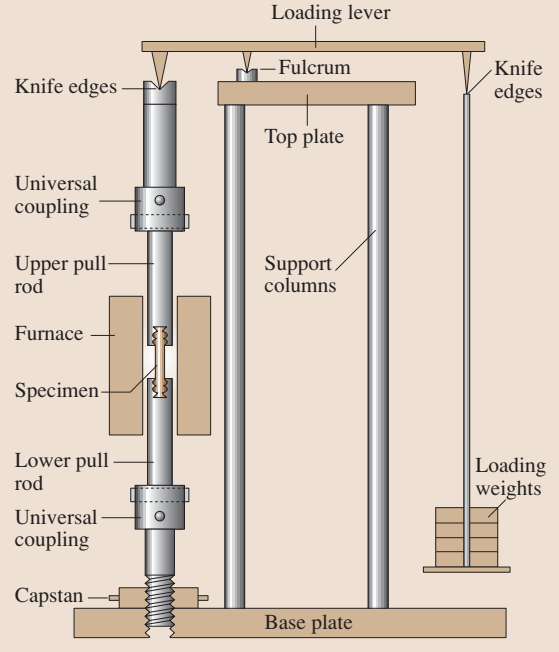


Fig. 3.60 Schematic of a constant-load creep-testing setup

der of 0.5 μm). Environmental effects can complicate creep tests by causing premature failures unrelated to elongation and thus must either mimic the actual service conditions or be controlled to isolate the failures to creep mechanisms. Uniformity of the applied stress is critical if the creep tests are to be interpreted properly. Figure 3.60 shows a typical creep testing setup.

The curve in Fig. 3.61 illustrates the idealized shape of a creep (strain–time) curve. The slope of this curve ($\mathrm{d}\varepsilon/\mathrm{d}t$ or $\dot{\varepsilon}$) is referred to as the creep rate. The initial strain $\varepsilon_\mathrm{i} = \sigma_\mathrm{i}/E$ is simply the elastic response to the applied load. The strain itself is usually calculated as the engineering strain, $\varepsilon = \Delta l/l_0$. The primary region (I) is characterized by transient creep with decreasing creep rate due to the creep resistance of the material increasing by virtue of material deformation. The secondary region (II) is characterized by steady-state creep (the creep strain rate $\dot{\varepsilon}_\mathrm{min} = \dot{\varepsilon}_\mathrm{ss}$ is constant) in which competing mechanisms of strain hardening and recovery may be present. The tertiary region (III) is characterized by increasing creep strain rate in which necking under constant load or consolidation of failure mechanism occur prior to failure of the test piece. The relative significance of the three creep stages depends on the temperature, creep stress, and material.

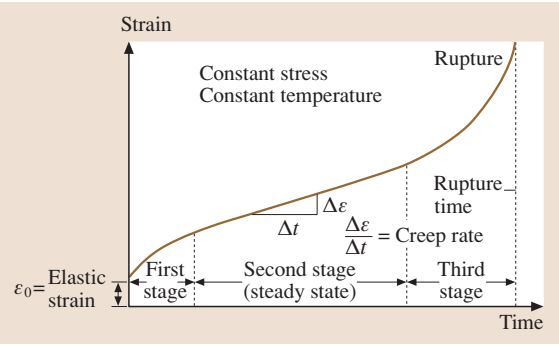


Fig. 3.61 An idealized creep curve showing the three stages during creep (after [3.54])

Traditionally, a creep curve is described by the following empirical relation, due to *Andrade* [3.54]:

$$\varepsilon = (1 + \beta t^{1/3}) \exp(\kappa t) \,, \tag{3.59}$$

where β and κ are empirically determined constants related to the primary and steady stages of the creep curve, respectively. However, for engineering applications, it is the steady-state creep rate that is of major concern; for example, what is the permissible stress needed to produce a minimum strain rate of 10^{-6}/h (i. e., strain of 0.01 in 10 000 h)? The *Mukherjee–Bird–Dorn* equation (3.60) is often used as a scaling relation

$$\frac{\dot{\varepsilon} k T}{D_0 G \boldsymbol{b}} = A_o \left(\frac{\sigma}{G}\right)^n \left(\frac{d}{\boldsymbol{b}}\right)^p \exp\left(-\frac{Q_c}{\mathrm{RT}}\right) \,, \tag{3.60}$$

where $\boldsymbol{b}$ is Burgers vector, d is grain size, D_o is the self-diffusion coefficient, G is the shear modulus, k is Boltzmann's constant, σ is the applied stress, Q_c is the activation energy for creep, R is the universal gas constant, T is the absolute temperature, and A, p, and n are dimensionless constants. Various creep mechanisms have been identified (both theoretically and experimentally) and are classified accordingly as:

1. Diffusion creep
2. Dislocation creep
3. Power-law breakdown

In diffusion creep processes atomic vacancies generated close to grain boundaries normal to the applied stress migrate to grain boundaries parallel to the applied stress, where they are absorbed. This process is leads to a shape change, but it does not involve dislocation flow (as the higher stress processes do). This diffusive transport of vacancies at higher temperatures occurs through the bulk of the grain (Nabarro–Herring creep) whilst at lower temperatures it occurs along the grain boundaries (Coble creep). Creep rates are reported

to be inversely proportional to the square of the grain size for Nabarro–Herring creep ($p = -2$ in (3.60)) and inversely proportional to the cube of the grain size for Coble creep ($p = -3$). When creep is controlled by diffusion alone, $n = 1$ in (3.60).

At higher stresses steady-state creep occurs by dislocation glide plus climb and values of n are typically 4–5; this is called the power-law regime. The upper boundary of the power-law creep regime is defined by the ratio of the applied shear stress to the elastic shear modulus that corresponds to the onset of general plasticity. For face-centered cubic (fcc) metals this ratio is given as 1.26×10^{-3} and deformation at stresses exceeding this value is said to be in the power-law breakdown regime. In the power-law breakdown regime, the Dorn equation is no longer valid. In the power-law and power-law breakdown regime, the steady-state creep rate is independent of grain size, i. e., $p = 0$. By solving the equations for diffusional flow, power-law creep, and general plasticity, it is possible to prepare deformation mechanism diagrams for a given material. The diagram for pure copper with a 100 μm grain size is reproduced in Fig. 3.62.

Two important phenomenon related to creep are *superplasticity* and *stress relaxation*. Superplasticity is the ability of a material to undergo large elongation without failure, see [3.56] for an overview. Stress jump tests are carried out to assess superplasticity of a material quickly at a given temperature by measuring the flow stress at different rates of loading, keeping the temperature constant. The strain-rate exponent, m ($= 1/n$, (3.59)), can then be evaluated and should be close to 0.5 for superplastic behavior to be observed. To determine the stress relaxation of a material, the specimen is deformed a given amount, and the decrease in stress is recorded over a prolonged period of exposure at a constant elevated temperature. The stress-relaxation rate is the slope of the curve at any point.

The goal in engineering design for creep is to predict performance over the long term. To this end, one of three approaches is applied:

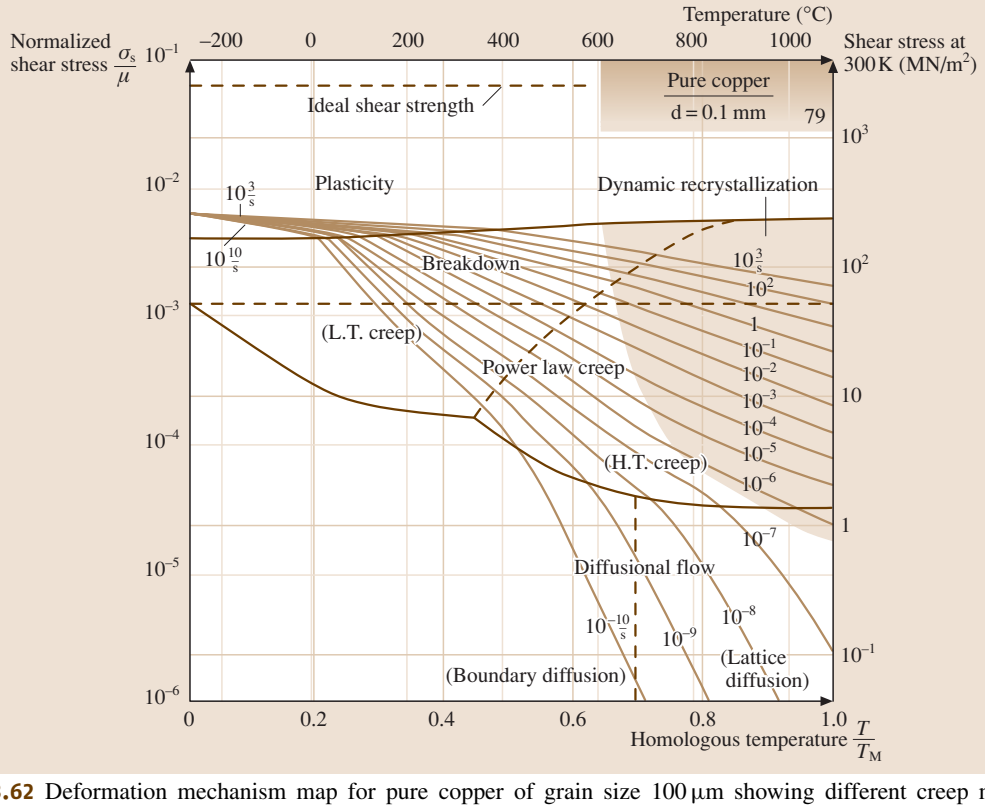


Fig. 3.62 Deformation mechanism map for pure copper of grain size 100 μm showing different creep mechanisms operating under a given temperature–stress region (after [3.55])

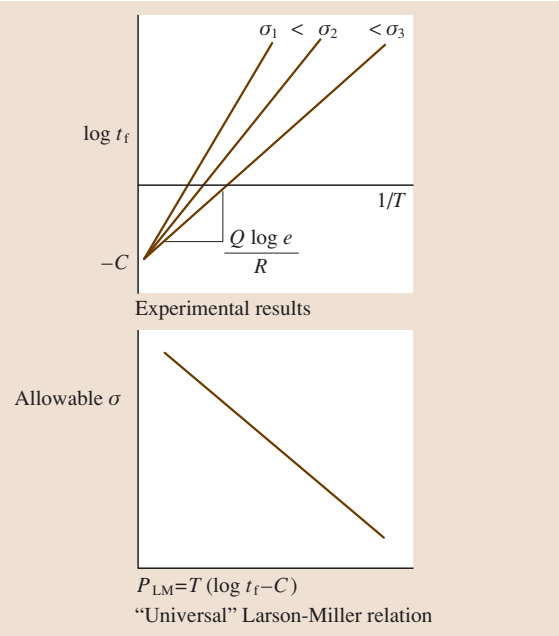


Fig. 3.63 Summary of the Larsen–Miller method for creep life prediction

- Stress-rupture tests. A large number of tests are run at various stresses and temperatures to develop plots of applied stress versus time to failure. While it is relatively easy to use these plots to provide estimates of stress rupture life, it is a very expensive and time consuming to develop these plots. Additionally, extrapolation of the data can be problematic.
- Minimum strain rate versus time to failure. This type of relation is based on the observation that strain is the macroscopic manifestation of the cumulative creep damage. A critical level of damage, independent of stress and/or temperature, is then defined as the failure criterion as follows:

$$\dot{\varepsilon}_{\min} t_f = C \propto \varepsilon_f \,. \qquad (3.61)$$

A log–log plot of $\dot{\varepsilon}_{\min}$ versus t_f or Monkman–Grant chart can then be constructed from a relatively few creep tests to determine the value of the empirical constant C and be used to predict creep life.

- Temperature-compensated time. In these methods, a higher temperature is used at the same stress so as to cause a shorter time to failure such that temperature is traded for time. In this form of accelerated testing it is assumed that the failure mechanism does not change and hence is not a function of temperature or time. In the most commonly used method, the Larson–Miller parameter P_{LM} at a given stress is expressed as

$$P_{LM} = 2.303 \frac{Q_c}{R} = T(C_1 + \log t_f) \,, \qquad (3.62)$$

where C_1 is the Larson–Miller constant, typically ranging between 25 and 60. Experimental data in terms of log t_f and $1/T$ at a given stress is plotted to estimate Q_c and C_1 as in Fig. 3.63.

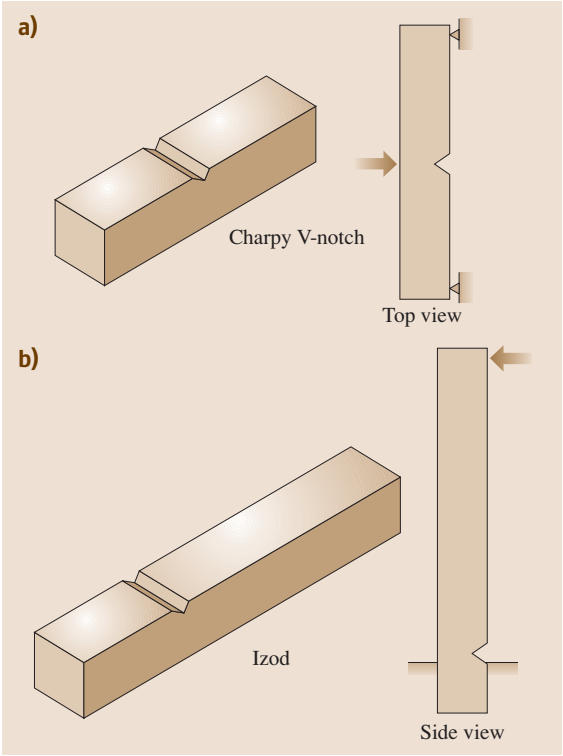


Fig. 3.64a,b Specimen geometry and test procedure for **(a)** Charpy V-notch test and **(b)** Izod impact test (after [3.13])

3.3.3 Dynamic Mechanical Properties

In structural applications, members are often subjected to varying load/stress over time either in the form of vibrations or high-energy impacts. It is important to have an understanding of the effect of such forces on structural integrity to avoid catastrophic failure by fracture.

Impact Testing

Toughness is a qualitative measure of a material's ability to absorb impact energy by undergoing plas-

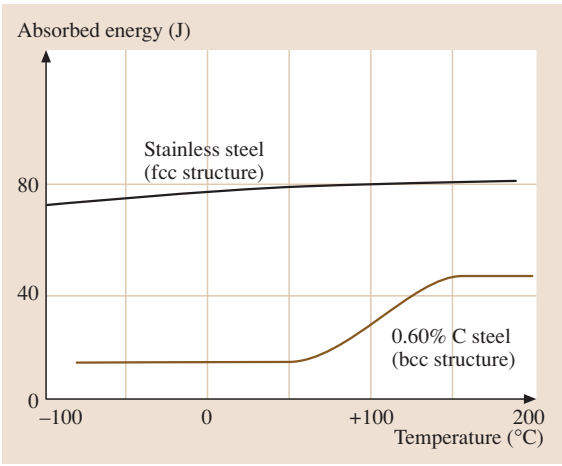


Fig. 3.65 Variation of absorbed energy as a function of temperature showing ductile to brittle transition (DBTT) in a plain carbon steel (after [3.13])

tic deformation. Notched-bar impact tests are generally used to detect the tendency of a material to fail in a brittle manner. Two classes of notched specimens are commonly used for these class of tests, namely the *Charpy* notched bar and the *Izod* specimen (Fig. 3.64). The specimen may have a square or circular cross-section and a V notch is machined either at the center (Charpy) or towards one end (Izod) of the specimen. The impact load is then applied by a heavy swinging hammer as indicated in Fig. 3.64. The presence of the notch creates a triaxial state of stress on the fracture plane. The response of the sample is usually measured by the energy absorbed in fracturing the specimen and can be estimated from the loss in kinetic energy before and after the impact.

The notched-bar impact test is most meaningful when conducted over a range of temperatures. Most metallic materials undergo a transition from a ductile to brittle mode of fracture with decreasing temperature. This transition temperature is called the ductile to brittle transition temperature (DBTT) and is an important design parameter. Figure 3.65 shows the typical variation of absorbed impact energy as a function of temperature for two steels, a ferritic (bcc structure) and an austenitic (fcc structure) one. The material with lowest DBTT should be preferred in structural applications at low temperatures to avoid catastrophic failure, hence, austenitic stainless steels are recommended.

Cyclic Testing

It is well known that a component subjected to a load well below its yield stress fails by fatigue, i. e., fluctuating load over a period of time (Fig. 3.66). Fatigue failure usually occurs without any obvious warning and is usually accompanied by fracture. A periodic stress cycle of the kind shown in Fig. 3.67a and b comprises a mean stress, $\sigma_m (= (\sigma_{max} + \sigma_{min})/2)$, and an alternating stress amplitude $\sigma_a (= (\sigma_{max} - \sigma_{min})/2)$. The stress ratio R is then defined as

$$R = \frac{\sigma_{min}}{\sigma_{max}} , \qquad (3.63)$$

where σ_{max} and σ_{min} are the maximum and minimum stress, respectively. The basic method of presenting engineering fatigue data is by means of the S–N curve, which represents the dependence of cycles to failure N on the maximum applied stress σ_{max}. Most investigations of fatigue properties are made by means of a rotating beam machine, where $\sigma_m = 0$ and $R = -1$. Figure 3.67 shows a schematic of the test apparatus and a typical S–N curve for this type of test. For most ferrous alloys, the S–N curve becomes horizontal at a certain limiting stress called the endurance or fatigue limit. Most nonferrous metals have S–N curves that do not show a true endurance limit and in such cases it is customary to define the endurance limit as the maximum stress that does not cause failure after 5×10^8

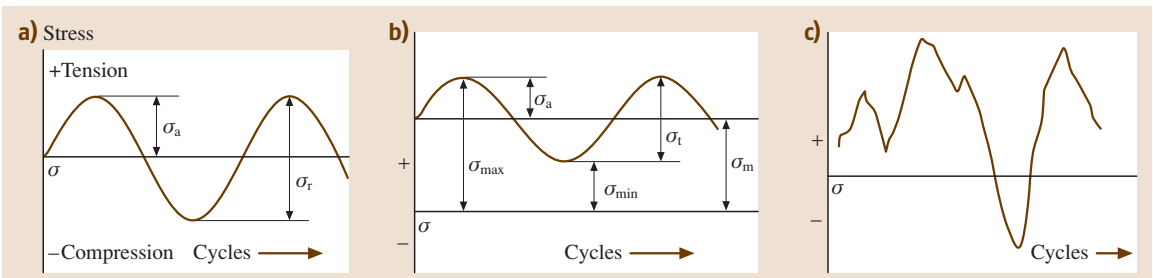


Fig. 3.66 **(a)** and **(b)** definition of different stress parameters during cyclic stress testing of materials. **(c)** A typical stress variation curve for an aeroplane foil in service (after [3.13])

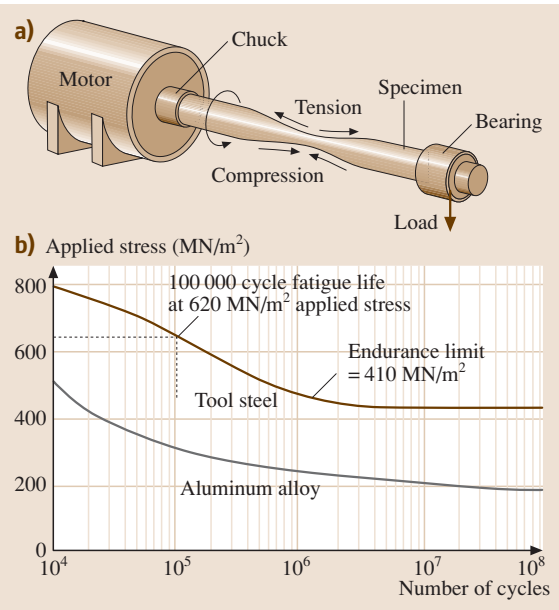


Fig. 3.67 **(a)** Schematic of a rotation bending fatigue testing equipment and **(b)** typical stress–number of cycles (S–N) curve for a ferrous and nonferrous alloy (after [3.54])

cycles. It must be noted that considerable scatter is observed during fatigue testing and it is standard to test three or four samples at a given stress level.

Application of a cyclic load, i. e., σ_a, leads to a cyclical strain response $\Delta\varepsilon_a$, which comprises elastic $\Delta\varepsilon_e$ and plastic $\Delta\varepsilon_p$ components. Figure 3.68 shows the variation of the elastic, plastic, and total strain amplitude as a function of number of cycles to failure. Based on this figure, it is easy to visualize that, below a particular *transition life* ($N_{f,t}$), plastic deformation controls the cycles to failure, while above the transition life elastic strain is the major source of fatigue damage. The two regions are termed low- or high-cycle fatigue, respectively, and the transition life is usually close to 10^4 cycles.

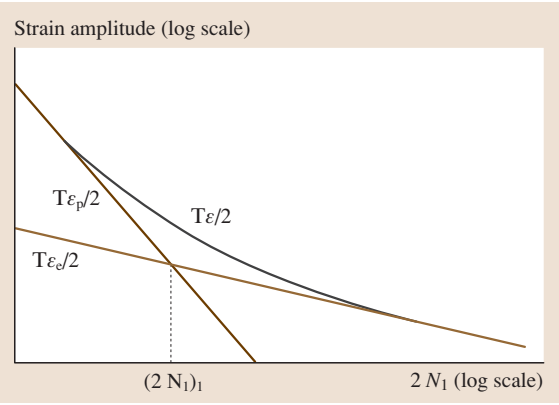


Fig. 3.68 Definition of a transition life between low- and high-cycle fatigue (after [3.57])

The fatigue process of a material can be divided into the following stages:

- *Crack initiation*, which including the early development of fatigue damage due to localization of slip at persistent slip bands (PSB) or embryonic cracks.
- *Slip-band crack growth*, which involves the deepening of the initial crack on planes of high shear stress. This is frequently called stage I crack growth.
- *Crack growth on planes of high tensile stress*, which involves the growth of well-defined cracks in a direction normal to the maximum tensile stress. This is usually called stage II crack growth.
- *Ultimate ductile failure*, which occurs when the crack reaches sufficient length that the remaining cross section cannot support the applied load.

In general, larger proportions of the total cycles to failure are involved with the propagation of stage II cracks in low-cycle fatigue than in high-cycle fatigue, while stage I crack growth comprises the largest segment for low-stress high-cycle fatigue. If the tensile stress is high, as in the fatigue of specimens with preexisting surface flaws or notches, stage I crack growth may not be observed at all. Specialized crack growth-rate tests using specimens which have been precracked in fatigue (see below) are employed to establish material selection criterion and to establish the effect of the following factors that are known to have a significant influence on fatigue life:

1. Stress or strain range
2. Mean stress
3. Surface finish and quality
4. Surface treatments
5. Load sequence and overload

A typical test results from these crack growth studies is shown in Fig. 3.69. Additional factors such as environmental conditions, elevated temperatures, and corrosive media drastically affect fatigue life and accelerate failure.

Fracture Mechanics

New nondestructive testing techniques allow designers to adopt a more *damage-tolerant* approach to structural

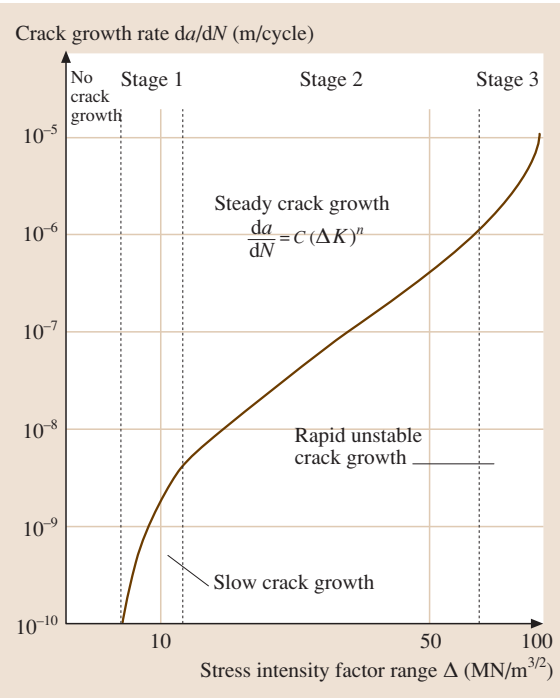


Fig. 3.69 A schematic fatigue crack-growth curve showing the three regimes of crack growth (after [3.54])

design. Accordingly, material with defects is no longer considered as failed but can be further used in service provided it is safe against *fast fracture*. The criterion for continued safe operation is that the strain energy release rate (called G for brittle materials and J for ductile materials) should be less than a critical value. The strain energy release rate is the amount of energy per unit length along the crack edge that is supplied by the elastic and plastic energy in the body and by the applied force in creating the new fracture surface area.

From linear elastic theory, the stress field around a crack (Fig. 3.70) can be expressed as

$$\lim_{r\to 0} \sigma_{ij}^M = \frac{K_M}{\sqrt{2\pi r}} f_{ij}^M(\theta) \,, \tag{3.64}$$

where the superscript and subscript "M" denotes the mode of load application (Fig. 3.71), f_{ij} is a function of location, and K is the stress intensity factor expressed as

$$K = Y\sigma\sqrt{\pi a} \,, \tag{3.65}$$

where Y is a dimensionless factor dependent on the sample geometry and a is the crack half-length. K can be related to G according to the following relations:

$$G = \begin{cases} \frac{K^2}{E} & \text{(plane stress)} \\ \frac{K^2}{E}(1-\nu^2) & \text{(plane strain)} \end{cases} \tag{3.66}$$

For ductile materials, the equivalent of energy release rate G is the J-integral, which is defined as

$$J = \int_\Gamma \left(W\,\mathrm{d}y - \overline{T}\frac{\partial \overline{u}}{\partial x}\,\mathrm{d}s \right) , \tag{3.67}$$

where W is the load per unit volume, Γ is the path of the line integral that encloses the crack tip, $\mathrm{d}s$ is the increment of the contour path, and $\overline{\boldsymbol{T}}$ and $\overline{\boldsymbol{u}}$ are the outward traction and normal vectors, respectively, on $\mathrm{d}s$. The first term corresponds to the elastic component while the second term corresponds to the plastic energy due to the crack.

As mentioned earlier, the stress intensity factor K should be less than a critical value (K_c) as a design

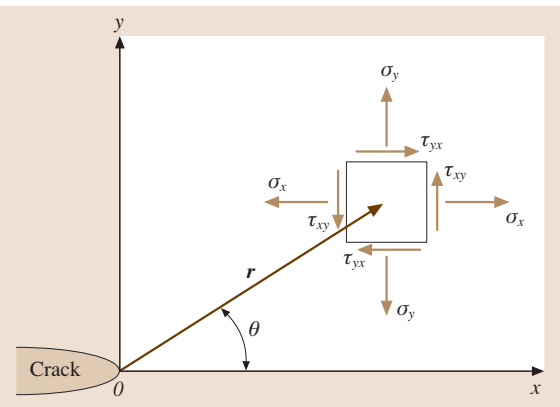


Fig. 3.70 Stress distribution at a point (r, θ) in the material stressed under a far-field tensile stress (mode 1) away from an elliptical crack (after [3.57])

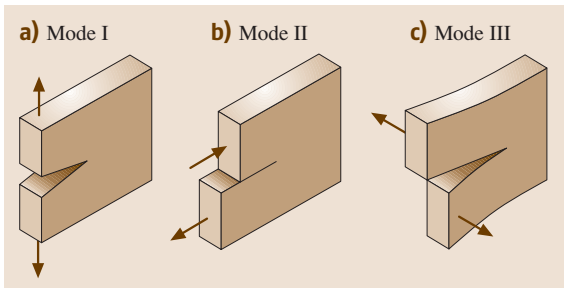


Fig. 3.71a–c Schematic representation of the three crack opening modes: **(a)** tensile, **(b)** shear, and **(c)** bending (after [3.53])

Table 3.3 Standards for mechanical testing of materials [3.58]

Test type	Standard
Methods of mechanical testing	ASTM E6–98
Tensile test of metallic materials	ASTM E8–98
Elevated-temperature tension tests for metallic materials	ASTM E21–92
Young's modulus, tangent modulus, and chord modulus	ASTM E111–97
Brinell hardness of metallic materials	ASTM E10–96
Vickers hardness of metallic materials	ASTM E92–82
Rockwell hardness and Rockwell superficial hardness of metallic materials	ASTM E18–97a
Hardness conversion tables for metals	ASTM E140–97
Microhardness of materials	ASTM E384–89
Compression testing of metallic materials at room temperature	ASTM E9–89a
Compression tests of metallic materials at elevated temperatures with conventional or rapid heating and strain rates	ASTM E209–65
Bend testing of mechanical flat materials for spring applications	ASTM E855–90
Shear modulus at room temperature	ASTM E143–87
Conducting creep, creep-rupture, and stress-rupture tests of metallic materials	ASTM E139–96
Stress relaxation for materials and structures	ASTM E328–86
Notched-bar impact testing of metallic materials	ASTM E23–96
Conducting force-controlled constant-amplitude axial fatigue tests of metallic materials	ASTM E466–96
Constant-amplitude low-cycle fatigue testing	ASTM E606–92
Measurement of fatigue crack growth rates	ASTM E647–95a
Measurement of fracture toughness	ASTM E1820–96

criterion. The critical stress intensity factor in tension mode, K_{Ic}, is a material property and can be interpreted as the inherent resistance of a material to failure. Hence, it is frequently called the *fracture toughness* of a material. Like other mechanical properties, it is determined experimentally as described below. The same test procedure applies to the determination of the J-integral, though the test data is treated differently. The test specimen may be a single-edge notched beam, com-

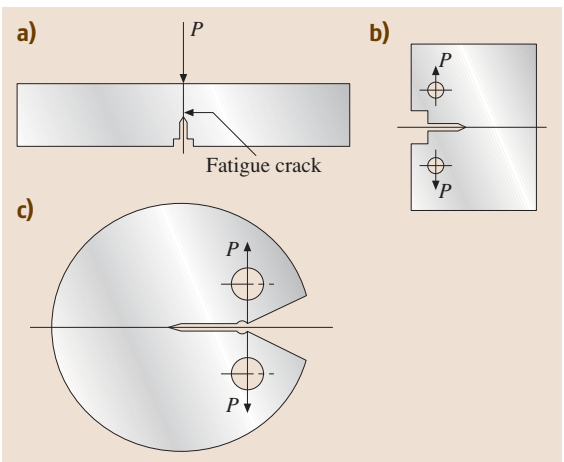


Fig. 3.72a–c Standard samples for fracture toughness (K or J) measurement: **(a)** single-edge bend, **(b)** compact tension (CT), and **(c)** cylindrical disc (after [3.53])

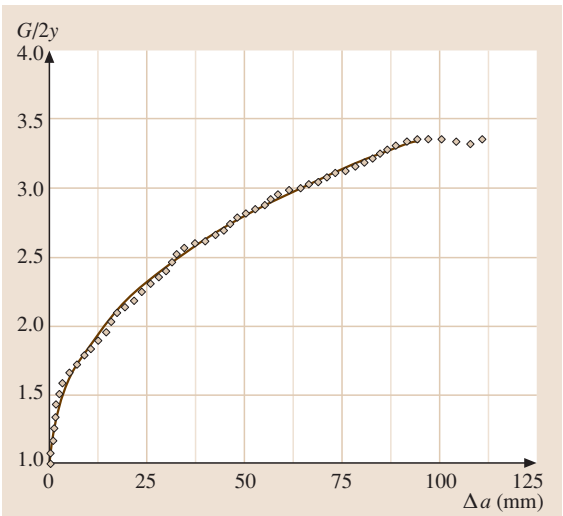


Fig. 3.73 A typical R-curve for a ferrous alloy showing the resistance to unstable crack extension (after [3.53])

pact tension or cylindrical disc (Fig. 3.72). Load line displacement is recorded as a function of applied load. A fatigue precracked test specimen is loaded in tension or bending to induce either:

1. Unstable crack extension (or *fracture instability*) or
2. Stable crack extension (or *stable tearing*)

The first method is used to determine the value of fracture toughness at the point of instability, while the second method results in a continuous relationship for fracture toughness versus crack extension (called the *R*-curve, Fig. 3.73). For *R*-curve determination, crack extension is also recorded simultaneously by optical or electrical means. The recorded data is then used to evaluate K_{Ic}, J_{Ic} or the J–R curve using standard relations. K_{Ic} is independent of the specimen geometry only under plain-strain conditions and this criterion should be assessed carefully. Similar crack growth tests may also be used to evaluate the performance of a material under creep and/or fatigue.

Table 3.3 summarizes standards for mechanical testing of materials according to ASTM [3.58].

3.4 Physical Properties

While the prime design criterion in most applications in mechanical engineering is mechanical properties (Sect. 3.3), physical properties are instead decisive for most applications as functional materials. As some of these materials are of paramount importance in fields related to mechanical engineering such as microelectronics, mechatronics, and the production, conversion, and distribution of electric power, we will briefly discuss in this section selected properties such as electrical and thermal conductivity with respect to materials in mechanical engineering, i. e., metals, ceramics, glasses, and polymers, as described in more detail in Sect. 3.6. Particularly, a discussion of the broad and still emerging fields of *magnetism and superconductivity* and *semiconducting materials* must be omitted here. For in-depth information, the interested reader is referred to the recent version of the *Encyclopedia of Magnetic and Superconducting Materials* [3.59] and to the *Springer Handbook of Condensed Matter and Materials Data* [3.1].

3.4.1 Electrical Properties

Ohm's Law and Electrical Conductivity

The relation between the voltage U (in Volts, V) and the current I (in Ampères, A) in an electric conductor (often in the form of a wire) is given by (the macroscopic) form of *Ohm's law* as

$$R = \frac{U}{I} , \tag{3.68}$$

where R is the resistance (in Ohms, Ω) of the material to the current flow and depends critically on the geometry and (intrinsic) properties of the material, therefore

$$R = \rho \frac{l}{A} = \frac{l}{\sigma A} , \tag{3.69}$$

where l is the length and A is the cross-section of the conductor; ρ (Ω m) and σ ($\Omega^{-1}\,\mathrm{m}^{-1}$) are the *electrical resistivity* and *electrical conductivity*, respectively, being specific for the material under consideration. Combining (3.68) and (3.69) yields

$$j = \frac{I}{A} = \sigma \frac{V}{l} = \sigma E , \tag{3.70}$$

with the current density j ($\mathrm{A/m^2}$) and the electric field strength E (V/m). Alternatively, j is given by the product of the number of charge carriers n, the charge of each carrier q, and the average *drift velocity* v of the carriers, thus

$$j = nqv . \tag{3.71}$$

Setting (3.70) and (3.71) equal yields the microscopic form of Ohm's law, which is more relevant for materials engineers

$$\sigma = nq \frac{v}{E} = nq\mu . \tag{3.72}$$

The term v/E is called the *mobility* μ ($\mathrm{m^2\,V^{-1}\,s^{-1}}$) of the charge carriers. While the charge q of the carriers of the electric current is a constant, one may readily recall from (3.72) that the electrical conductivity of materials can be controlled essentially by two factors, namely:

1. The number of charge carriers n
2. Their mobility μ

While electrons are the charge carriers in conductors (metals), semiconductors, and many insulators, ions carry the charge in ionic compounds. Therefore, in pure materials the mobility μ depends critically on the bonding strength and – in addition in ionic compounds – on

Table 3.4 Electrical conductivity σ at RT for selected (pure) materials [3.1, 54]

Material	Electric conductivity σ ($\Omega^{-1}\,m^{-1}$)
Al	3.77×10^{7}
Ag	6.80×10^{7}
Au	4.26×10^{7}
Cu	5.98×10^{7}
Fe	1.00×10^{7}
Mg	2.257
Ni	1.46×10^{7}
Pb	5.21×10^{6}
Ti	2.56×10^{6}
W	1.82×10^{7}
Zn	1.84×10^{7}
Si	5×10^{-4}
Polyethylene	10^{-13}
Polystyrene	10^{-15}–10^{-17}
Al_2O_3	10^{-12}
Diamond	$< 10^{-16}$
SiC	1–10
SiO_2 (silica)	10^{-15}

diffusion rates, which in turn gives rise to the tremendous variation of electrical conductivity over more than 20 decades, see Table 3.4.

Effect of Temperature on the Electrical Conductivity of Metallic Materials

When heat is applied to metallic materials the atoms gain thermal energy and vibrate at a particular amplitude and frequency. Thus, increasing the temperature increases the probability of scattering electrons within the crystal, which ultimately leads to a reduction of the mobility of electrons; the resistivity ρ_T (of a pure material) at a particular temperature changes according to

$$\rho_T = \rho_{RT}[1 + a(T - RT)] , \tag{3.73}$$

where a is the *temperature resistivity coefficient* and "RT" indicates room temperature. The relationship between resistivity and temperature is linear over a wide temperature range; values of a for metals are positive and are tabulated in [3.1].

Effect of Lattice Defects on the Electrical Conductivity of Metallic Materials

An additional contribution to electron scattering stems from all kinds of lattice imperfections, as listed in Sect. 3.1.2. As a representative, the increase in resistivity due to atoms in solid solution can be described as

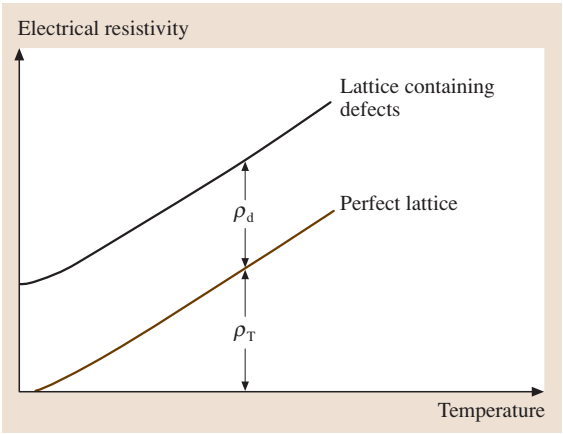


Fig. 3.74 Dependence of electrical resistivity of metallic materials on temperature; for further explanations see text

$$\rho_d = C(1 - x)x , \tag{3.74}$$

where ρ_d is the increase in resistivity due to the lattice defects present in the material and x is defined as the molar fraction of these defects (Sect. 3.1.2); C is the defect resistivity coefficient. Thus, the overall resistivity is

$$\rho = \rho_T + \rho_d . \tag{3.75}$$

Note that the effect of defects is virtually independent of temperature (Fig. 3.74).

3.4.2 Thermal Properties

As outlined in Sect. 3.1.2 the atoms in a material have a minimum free energy at absolute zero. However, as mentioned in the previous subsection supply of thermal energy causes the atoms to vibrate at a particular amplitude and frequency. This gives rise to a number of physical effects and related quantities such as the *heat capacity* or *specific heat*, *thermal expansion*, and *thermal conductivity*, which will be discussed briefly in the following subsections.

Heat Capacity and Specific Heat

Since vibrations of atoms are transferred through the whole crystal as elastic waves, heating up or cooling down of a material is realized by accepting or loosing *phonons* of energy

$$E = \frac{hc}{\lambda} = h\nu . \tag{3.76}$$

Table 3.5 Specific heat c of selected materials at RT [3.1, 54]

Material (metals)	c ($J\,kg^{-1}\,K^{-1}$)	Material (others)	c ($J\,kg^{-1}\,K^{-1}$)
Al	900	Al_2O_3	837
Cu	385	Diamond	519
Fe	444	SiC	1047
Mg	1017	Si_3N_4	712
Ni	444	SiO_2 (silica)	1109
Pb	159	Polyamide	1674
Ti	523	Polystyrene	1172
W	134	Water	4186
Zn	389	Nitrogen	1042

Then, the *heat capacity* is the energy required to raise the temperature of *one mole* of a given material by one *Kelvin* (K). It can be determined by various methods either at constant pressure C_p or at constant volume C_V. As depicted in Fig. 3.60 the heat capacity approaches a nearly constant value of $C_p = 3R \approx 25\,J\,mol^{-1}\,K^{-1}$ at sufficiently high temperatures for most metallic (above $\approx$ RT) and ceramic (> 800 K) materials. An exception to this rule is the (toxic) element Be which has a mere $C_p \approx 16\,J\,mol^{-1}K^{-1}$ [3.1].

By contrast, the *specific heat* c is the energy required to raise the temperature of a particular weight or mass of a material by 1 K. The relationship between heat capacity and specific heat is simply given by $c = C_p/M$, where M is the atomic mass (see periodic table). For engineering applications, specific heat is more appropriate to use than heat capacity. A compilation of the specific heat of typical materials is given in Table 3.5. Data on water (liquid) and nitrogen (gas) are given also for comparison, with H_2O having the highest value of specific heat. Note that neither specific heat nor heat capacity depend significantly on microstructure.

Table 3.6 Linear coefficient of thermal expansion at RT for selected materials [3.1, 54]

Material	Linear coefficient of thermal expansion α ($10^{-6}\,K^{-1}$)
Al	23.03
Cu	16.5
Fe	12.3
Mg	26.1
Ni	13.3
Pb	29.1
Ti	8.35
W	4.31
Zn	25.0
0.2% C steel	12.0
304 stainless steel	17.3
Invar alloy (Fe-36%Ni)	1.54
Polyamide	80
Polystyrene	70
Al_2O_3	6.7
Diamond	1.06
SiC	4.3
Si_3N_4	3.3
SiO_2 (silica)	0.55

Thermal Expansion

As pointed out in Sect. 3.1.1 the lattice constant of a material is a measure of the strength of atomic bonding, which is in turn the result of force equilibrium between an attractive and a repulsive potential. If a material gains thermal energy, however, this equilibrium separation increases since the material is lifted from its energy minimum into a higher-energy state. The change in the dimensions of the material is usually measured by *dilatometry* as

$$\alpha = \frac{l_f - l_0}{l_0(T_f - T_0)} = \frac{\Delta l}{l_0 \Delta T}\,, \qquad (3.77)$$

where the indices "f" and "0" denote the final and initial values of length l and temperature T. Linear coefficients of thermal expansion α at RT for selected materials are listed in Table 3.6.

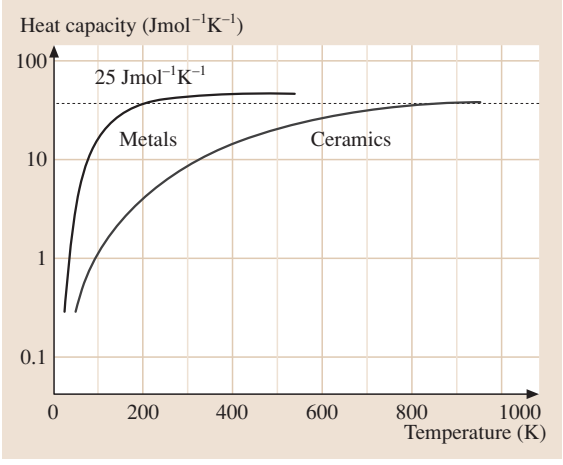


Fig. 3.75 Heat capacity as a function of temperature for metals and ceramics

Table 3.7 Thermal conductivity k of selected materials at RT [3.1, 54]

Material (metals/alloys)	k (W m^{-1} K^{-1})	Material (others)	k (W m^{-1} K^{-1})
Al	238	Al_2O_3	16
Cu	1401	Diamond	2320
Fe	80	Graphite	335
Mg	100	SiC	88
Ni	444	Si_3N_4	15
Pb	35	SiO_2 (silica)	1.34
Ti	22	ZrO_2	5.0
W	172	Polyamide	0.25
Zn	117	Polyethylene	0.33
0.2% C steel	100	Polyimide	0.21
304 Stainless steel	30		
Grey cast iron	80		
Cu-30% Ni	50		

Two conclusions can be drawn from the compilation in Table 3.6, namely that materials possessing strong atomic bonds, in particular covalently bonded materials such as many ceramics, have:

1. Low α values and
2. High melting points T_m

The latter relationship is shown for metals in Fig. 3.76.

A particular behavior must be noted for the Invar alloy Fe–36%Ni, which reveals that interaction with magnetic domains may suppress thermal expansion nearly completely until the Curie temperature is reached. This makes Invar attractive for bimetallic applications.

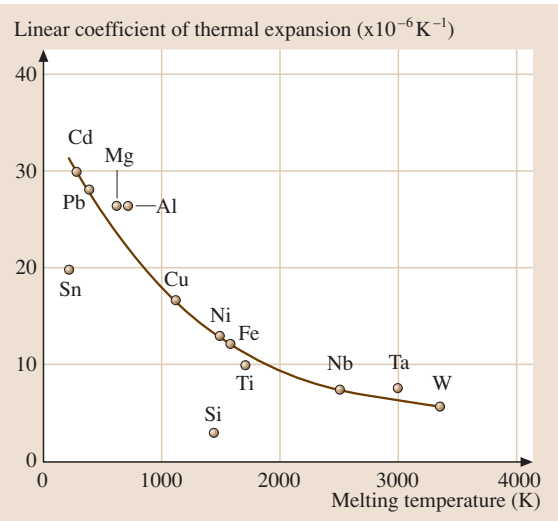


Fig. 3.76 Relationship between the linear coefficient of thermal expansion (at RT) and the melting point in metals

Thermal Conductivity

In essence, *thermal energy* is transferred in solid materials by two mechanisms:

1. Transfer of free (valence) electrons
2. Lattice vibrations (phonons)

of which the latter is closely related to the phenomenon of storing thermal energy, i. e., the *heat capacity* (see above).

Hence, the *thermal conductivity* k is a measure of the rate at which heat is transferred through the material and follows the relationship

$$\frac{Q}{A} = k\frac{\Delta T}{\Delta x}\,, \tag{3.78}$$

where Q is the heat transferred through a cross-section A induced by a temperature gradient $\Delta T/\Delta x$. Note the

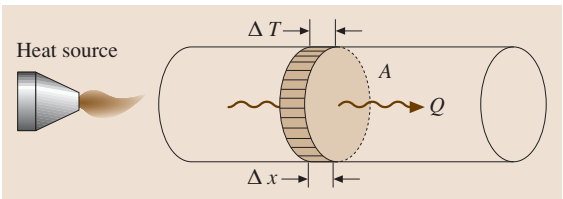


Fig. 3.77 Schematic of the method for measuring the thermal conductivity k according to (3.75, 78)

striking similarity between k and the diffusion coefficient D in mass transfer (3.14), where the *heat flux* Q/A is analogous to the flux of atoms j_D. A schematic experimental setup for measuring k is shown in Fig. 3.77, where heat is introduced on one side of a bar- or disc-shaped sample through a heat source and the change of temperature on the other side is measured as a function of time. The commonly employed technique is called the *laser flash method*.

Values for the thermal conductivity k of selected materials are listed in Table 3.7. A comparison yields that the k values of metals and alloys are usually much larger than those of ceramics, glasses, and polymers. This is due to the fact that in metals and alloys thermal energy is transferred through the movement of (loosely bonded) valence electrons which can be excited with little thermal energy into the conduction band. This leads to a relationship between thermal and electrical conductivity in many metals of the form

$$\frac{k}{\sigma T} = L = 2.3 \times 10^{-8}\,\mathrm{W\,\Omega\,K^{-2}}\,, \tag{3.79}$$

where L is the Lorentz constant.

In contrast, the prime energy transfer mechanism in ceramics, glasses, and polymers is vibration of lattices and (silicate or molecular polymeric) chains, respectively. Since the electronic contribution is absent, the thermal conductivity in these material classes is usually much lower than that in metals and alloys. An exception to the rule is *carbon* in its covalently bonded form as *diamond*, which has the highest k value and therefore commonly serves as a *heat sink* material.

The situation is reversed when the temperature of the materials is increased: the greater lattice and chain vibrations usually lead to an increase of the thermal conductivity in ceramics, glasses, and polymers. In metals and alloys the same mechanism applies in principle, however, the electronic contribution will be lowered, even though the number of carriers is increased, as their mobility is more strongly reduced due to increasing scattering effects. Therefore, thermal conductivity in metals and alloys usually decreases with increasing temperature. Like the electrical conductivity, thermal conductivity in metals and alloys also decreases with increasing number of lattice defects of various dimensionality (Sect. 3.7.2), introduced into the microstructure due to the increased electron scattering.

3.5 Nondestructive Inspection (NDI)

Nondestructive inspection (NDI) includes all methods to characterize a material without indenting, extracting samples, reducing its service capabilities or even destroying it. NDI includes defect detection and quantification, called nondestructive testing (NDT), and the assessment of material properties, called nondestructive evaluation (NDE). NDI is an integral part of component design, manufacturing, maintenance, and recycling of components.

More and more components are designed following the rule of fitness-for-service. This concept assumes the presence of a maximum undetectable-by-NDI defect. The design has to make sure that this defect does not become critical during a well-defined service period. To keep the safety coefficient at a predefined level the component will be larger or heavier than it should be without the defect. With increasing capabilities of NDI this maximum undetectable defect decreases, allowing the designer to reduce the component weight while keeping the safety coefficient at the same level.

In manufacturing, NDI enables the inspection of the whole output while destructive methods rely on a more or less satisfying quantity of samples being more or less representative for the current party. Besides suitability, the inspection speed is the deciding criterion for NDI application.

In maintenance there is no alternative to NDI. According to considerations of fracture mechanics the concept of damage tolerance requires the detection and characterization of all defects starting from an individually defined level. Depending on the findings of NDI the next service period may be shorter or longer. The typical requirement for inspection is a high probability of defect detection accompanied by a tolerable rate of false indications. Modern maintenance concepts include online monitoring of the structural health of a component or the whole construction.

All industrial branches use NDI, the best known being flying structures. However, pipelines, heat exchangers, vessels, bridges, and car components are also inspected nondestructively. We will focus on the most important and widely used methods in mechanical engineering but also touch on the promising field of structural health monitoring (SHM).

3.5.1 Principle of Nondestructive Inspection

The basic principle of NDI is shown in Fig. 3.78. The goal is either to detect relevant defects or to estimate quality parameters such as hardness, heat treatment or coating layers. This goal cannot directly be reached nondestructively. The only possible way is to measure physical properties such as conductivity, sound propagation or magnetic behavior. Both the quality parameters and the physical properties are defined by the material's structure. The challenge is to find a correlation between them. It is a matter of current research to complete the knowledge about the relations between the quality parameters to be inspected and the physical properties recordable nondestructively.

A wide variety of NDI sensors and transducers are known to record the physical properties locally or integrally. These signals are processed in an instrument and displayed in different forms. Some physical properties may be recorded by matrix sensors, immediately providing an image, while other properties have to be measured by point-like sensors that are hand-guided by an operator or mechanically guided by a scanner.

The indication is either an image, a vector or a scalar value. In most applications a threshold is used to separate appropriate from inappropriate quality (go/no go). For this purpose, differential measurements are most suitable. The current measurement is compared with the measurement of a master piece or the measurement of a neighboring area of the same object. If the difference is below the threshold, the object passes, otherwise it is failed. If quantitative assessment is required, calibration curves have to be taken basing on well-known samples with gradual variation of the parameter to be evaluated. Here, the increasing performance of numerical modeling is providing valuable help. In some cases NDI is expected to provide absolute values of a physical property, which is provided by dedicated instruments.

How reliable is NDI? To date no method is known to detect defects with a probability of 100%. Vice versa, all methods may produce false indications, e.g., indicate a defect in sound material. All NDI applications have to be optimized regarding both probabilities.

The following sections present a selection of approved methods and show the direction of future development. The references [3.60–64] are the most appropriate introductions to NDI. Further references on individual methods in each section provide more detailed information but cannot exempt the user from contacting experienced specialists. International standards and rules of application exist for all methods.

3.5.2 Acoustic Methods

Acoustic methods rely on the propagation of sound waves though the material. These waves may be excited by external or internal sources. In solids, longitudinal (compressional) and transverse (shear) waves may spread and are partially reflected and mode converted at boundaries. Additional wave modes may exist at the material's surface and in thin plates (Rayleigh and Lamb waves). While propagating through the material the waves are attenuated depending on the material's properties and the wave's frequency.

Ultrasonic Methods

Principle. The basic idea is to transmit a short elastomechanical wave packet into the material. If its wavelength is short enough it will interact with defects starting from approximately 0.5 mm in diameter. Therefore exiting frequencies from 0.5 to 15 MHz are required (ultrasonic frequencies). Depending on the material's attenuation the wave packet travels long distances through the material. In the transmission technique the pulse is received at the opposite side of the object whereas in the pulse-echo technique the reflected waves are recorded

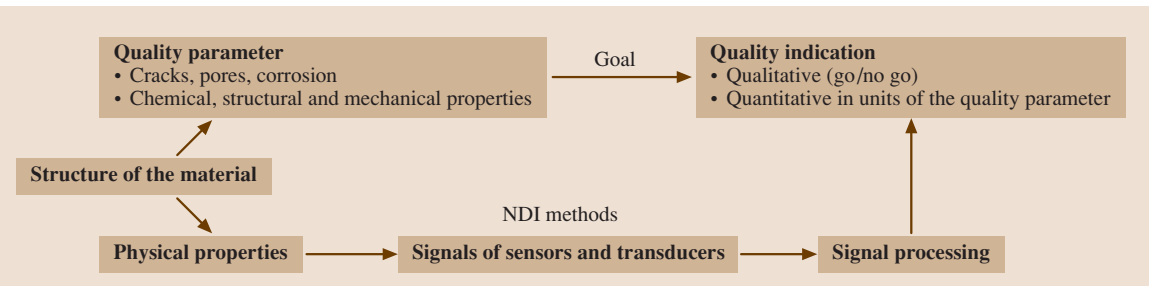


Fig. 3.78 The principle of nondestructive inspection

at the same side where they are fed. Nearly all metals, ceramics, concrete, and low-damping plastics and composites may be inspected [3.65, 66].

Probes (Transducers). For perpendicular wave propagation, perpendicular-incidence probes are attached to the surface. For nonperpendicular wave propagation angle-beam probes are required, commonly generating longitudinal waves that are diffracted into a longitudinal and a transverse wave at different angles at the material's surface. The acoustic waves must be coupled to the material using a couplant such as water or grease. Figure 3.79 presents a perpendicular-incidence probe consisting of a piezoelectric ceramic disc emitting and receiving the ultrasonic waves. A resin damping block absorbs the waves emitted in the reverse direction. The ultrasonic waves are passed through a protecting layer and the couplant into the material. After entering the material the sound field is characterized by local maxima and minima due to the interference of waves emitted from different parts of the piezoelectric disc. This initial area of the sound field is called the near field, where no inspection is possible. The near field becomes taller and ends in a final maximum of the sound pressure. From this point onwards the sound field diverges and the sound pressure gradually decreases. This area is best suited for NDI.

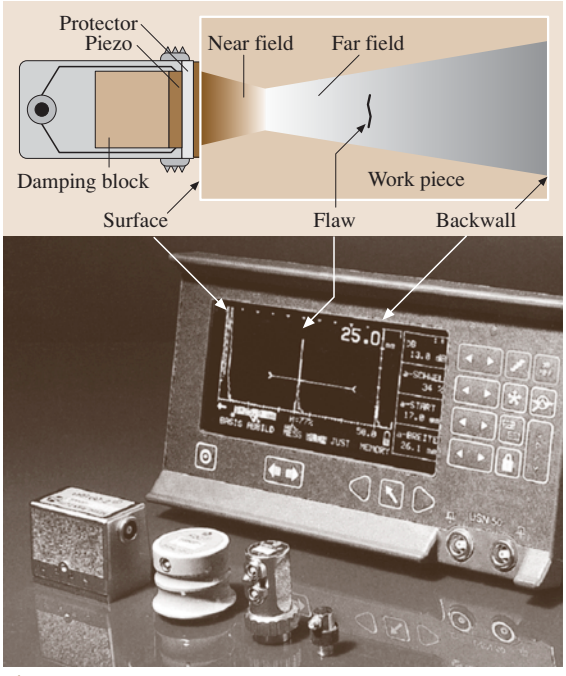


Fig. 3.79 Principle of ultrasonic inspection, instrument, transducers, and signal representation

NDT. Defects may be detected if they interact with the ultrasonic waves. The best detection performance is available for defects oriented perpendicularly to the sound field axis. Both transmission and pulse-echo techniques are suitable. With decreasing wavelength (increasing frequency) smaller defects may be detected but reflections from grain boundaries are also superimposed the signal. In Fig. 3.79 the screen of an ultrasonic instrument displays the A-scan formed by the rectified and amplified echo sequence. At the entrance a significant part of the sound energy is directly reflected back to the probe, producing the large entrance echo. At the back wall nearly all the sound energy is reflected. A properly oriented planar or volumetric defect produces an intermediate echo between the entrance and back wall echo. This echo may be analyzed to evaluate the defect's position and dimension. In weld inspection angle beam probes must mostly be used due to the nonsmooth surface of the weld and the orientation of possible defects.

NDE. The sound velocity and attenuation carry information about the structural and geometrical parameters of the material. The time-of-flight method is used for measuring the dynamic Young's modulus and for evaluating the structure of cast iron with spherical graphite, stress and strain assessment in steel, and the thickness of metal or nonmetal walls. Grain boundary reflections are welcome to estimate grain size or surface hardening depth. Sound attenuation measurement is also applied for structural characterization such as grain sizing and deformation-induced alterations.

Tendency. Instead of single transducers, one- or two-dimensional arrays are used two incline and shape the ultrasonic beam. Electronic movement enables fast imaging techniques. To avoid liquid couplants current research is focused on air-coupled techniques and thermally induced sound waves and interferometric read out.

Resonance Methods

Principle. The object is excited by a mechanical pulse or a continuous wave in a defined audible or ultrasonic range. It starts vibrating on its eigenmodes and the eigenfrequencies are recorded. The signal is analyzed in the time and frequency range, comparing the signals to those of one or more master pieces with well-known properties. Nearly all metals, ceramics, plastics,

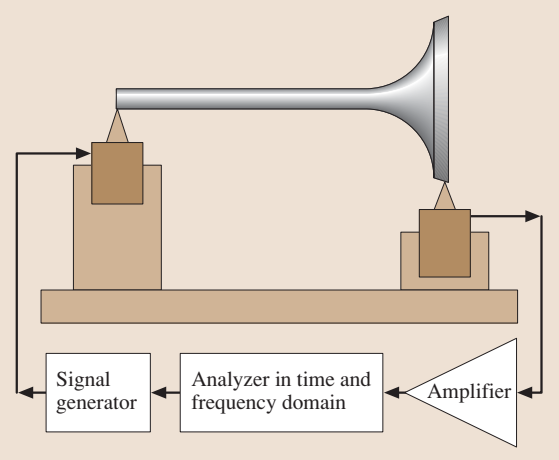


Fig. 3.80 Schematic view of valve inspection using resonance method (after [3.67])

and composites, and often the adhesive bonding between them, may be inspected according to the go/no go principle.

Setup and Probes. The object is positioned on a low-damping fixture and is excited by defined mechanical pulses using impact hammers, piezoactuators or electromechanical shakers. In the low-frequency range the acoustic response can be heard or is picked up by microphones via air coupling. At higher frequencies and lower amplitudes piezoelectric sensors are attached to the object or optical interferometers record the oscillation at one or more positions. Figure 3.80 shows the setup for valve inspection. Figure 3.81 presents the signals in the time and frequency domains.

NDT. Complex-shaped objects such as gears and cast housings are inspected for missing components, imperfect shape, cracks, and cavities. Mostly the frequency content of the response signal is analyzed. In a first step a representative amount of sound and flawed objects is analyzed in a broad frequency band. Comparing the response spectra, the most sensitive narrow bands are selected for defect detection.

NDE. The resonance method allows the estimation of structural damping, the comparison of elastic properties between identically shaped samples, and even the measurement of the dynamic Young's modulus with simply shaped specimens.

Tendency. With increasing sensitivity smaller defects become detectable. Extensive signal processing allows

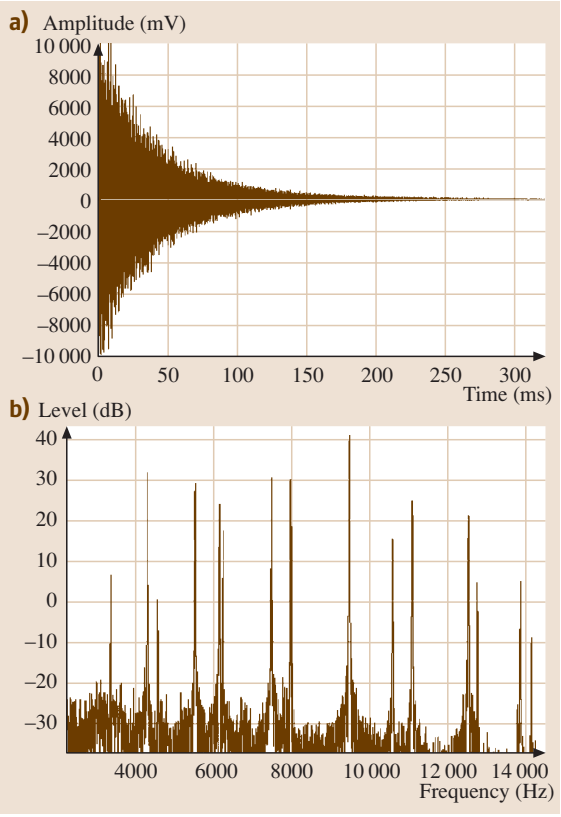


Fig. 3.81a,b Resonance analysis of valve: **(a)** signal in the time domain, **(b)** signal in the frequency domain (after [3.68])

the suppression of disturbing signals from the environment.

Acoustic Emission Analysis

Principle. When an object is loaded, defects grow and discontinuously radiate elastic wave bursts (Figs. 3.82, 3.83). These bursts are picked up and analyzed according to their spectral content, signal energy, and other specific parameters. Defect location becomes possible using time-of-flight differences to different sensors [3.69]. For leakage detection the continuous acoustic radiation arising at the leak point is recorded. For leakage location, signals from two or more separated sensors are correlated to define the time-of-flight difference.

Setup and Probes. For loading the object may be heated or stressed by different means. Piezoelectric probes with internal or closely attached external preamplifiers pick

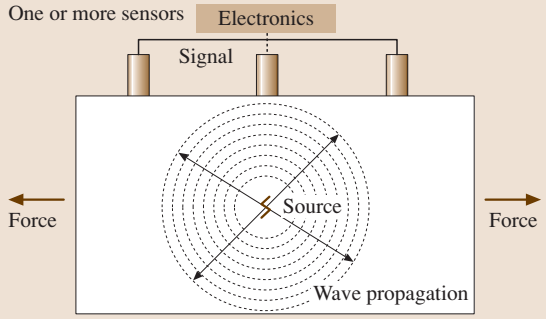


Fig. 3.82 Acoustic emission generated by a discontinuously growing crack

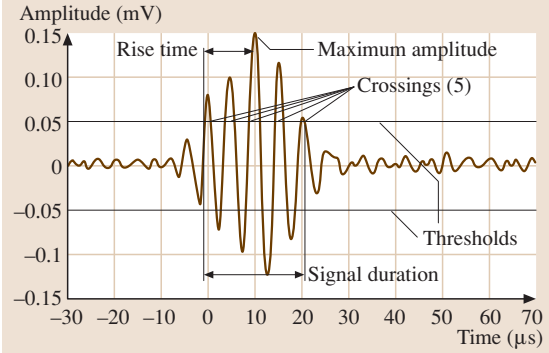


Fig. 3.83 Burst signal recorded by attached sensors

up the emission. Mostly, more than one sensor is required to cover the whole object or to locate the defects.

NDT. Growing cracks, hydrogen embrittlement, stress corrosion, and creep can generate acoustic emissions. The advantage of their analysis is that they enable information about the whole object to be attained at once. Inspection is commonly performed before running the object and during inspection breaks. Generally, acoustic emission may be recorded online, thus enabling structural health monitoring (Sect. 3.5.9). No information about defect dimensions is available.

Tendency. With increasing knowledge about signal generation, signal conversion, and nonresonant sensors more detailed information can be obtained by characterizing the source of radiation, thus making acoustic emission analysis more reliable.

3.5.3 Potential Drop Method

Principle

Once found, a surface crack's depth should often be estimated. For this, an electric current is passed perpendicularly across the crack. The current will be deflected by the crack, depending on its depth and length [3.70]. Two electrodes are contacted on both sides of the crack to measure the potential drop. Assuming that the crack is much longer than it is deep, the voltage primarily depends on the crack depth. Figure 3.84 shows the reason for increasing potential drop with increasing crack depth.

Probes

Modern probes combine the current supply and potential drop electrodes into one probe. Figure 3.85 presents

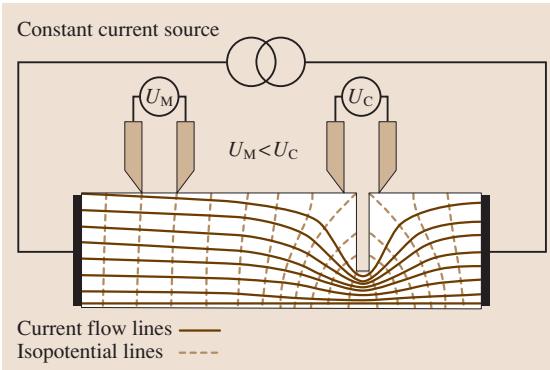


Fig. 3.84 Potential drop measurement of crack depth

Fig. 3.85 Potential probes are able to measure crack depth and inclination

such a four-point probe. The outer electrodes feed the current and the inner ones measure the voltage. For online observation of crack growth, electrodes may be permanently fixed to the object.

NDT

For crack depth assessment in ferrous steels alternating current is advantageous due to the skin effect. This effect causes a current concentration at near-surface regions while direct current spreads out much deeper. That is why the crack's influence on lengthening the current path is more pronounced with alternating than it is with direct current. After careful calibration the crack depth can be estimated with an accuracy of some tenths of a millimeter, taking into account that the first electric bridge between the crack faces defines the measured crack depth. To estimate crack inclination an additional electrode fixed a greater distance from the crack is necessary.

NDE

Potential drop measurement can be used for measuring the conductivity of metals. The advantage over the eddy-current method (Sect. 3.5.5) is its suitability also for ferrous steels, while eddy-current conductivity measurement may only be applied for nonferromagnetic materials. The disadvantage is the unavoidable direct contact to the metal, requiring at least local stripping of paint and corrosion products.

Tendency

Calibration curves of most common materials are stored in the instrument. Combined inductive feed and contact gauging is the subject of current investigations.

3.5.4 Magnetic Methods

Magnetic methods use the ferromagnetism of ferritic steels. A magnetic flux passed through the material orients the magnetic domains of the material, thus increasing the flux density. This orientation process is nonlinear and follows a hysteresis loop. When all domains are oriented according to the exciting field, the material is magnetically saturated. For NDI a number of physical properties may be used such as saturation induction, remanence, coercive force, and magnetic permeability [3.71].

Stray Flux and Magnetic Particle Inspection

Principle. If the magnetic flux passed through the ferromagnetic material faces a boundary to a less permeable area (e.g., air in a crack) it is refracted into this area nearly perpendicularly to the boundary. Figure 3.86 details this situation where a part of the flux spreads over the boundaries of the workpiece. This stray or leakage flux is much wider than the crack. The remaining flux

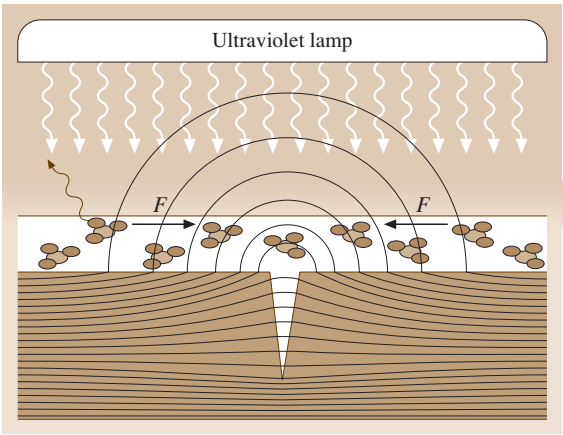


Fig. 3.86 Magnetic stray flux generates the force F attracting the magnetic particles

lines pass below the crack or cross it. They are not accessible for NDI. The estimation of crack depth is not possible.

Setup and Probes. The magnetic flux must be oriented perpendicularly to the crack. This flux may be excited, whether by permanent magnets, electric current or coils (electromagnets). For stationary equipment combined electric and electromagnetic excitation is preferred to produce a circular magnetic field. The objects are placed in the gap of a yoke that carries the magnetic flux as well as the electric current. Mobile excitation is possible by electromagnetic hand yokes and current electrodes. The stray flux can be detected by magnetic sensors such as flux gates, magnetoresistors, Hall sensors or moving coils, or even visualized by magnetic particles (magnetic particle inspection, MPI). After inspection the object has to be demagnetized.

NDT by Flux Sensors. In automatic inspection lines objects such as pipes, rods or sheets are moved through a magnetizing yoke. Between the poles sensors or sensor arrays are guided over the surface to detect and quantify the stray flux.

NDT by MPI. A suspension of high-permeability fluorescent powder in a low-viscosity carrier liquid is flushed over the surface. The magnetic particles are attracted by the stray flux, dragging the fluorescent particles with them. In a darkroom under ultraviolet illumination these particles become visible, indicating the crack (Fig. 3.87) [3.72]. For documentation photographs can be taken. In difficult conditions such as underwater in-

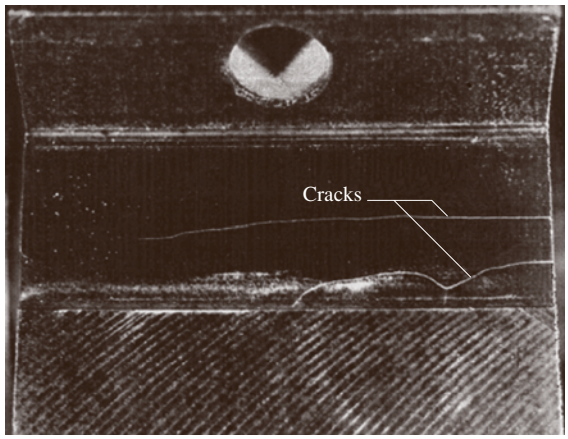


Fig. 3.87 Magnetic particle indication of cracks in a switch shaft

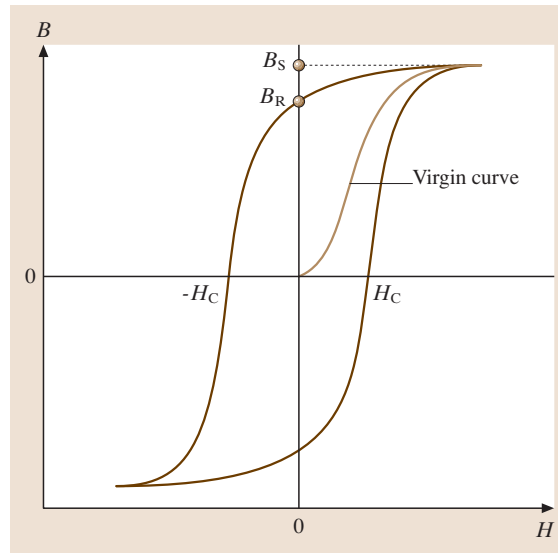


Fig. 3.88 Ferromagnetic hysteresis loop. B_S – saturation induction, B_R – remanent induction remaining on the part after removing the exciting magnetic field H, H_C – coercivity

spection the suspension is contained in a double-wall package. After exposing the package to the stray flux the suspension can be cured by a second component to fix the particles at their positions and to analyze the image later.

Tendency. To increase the reliability of MPI attempts are being made to record the MPI image using video cameras and process the image automatically. Instead of magnetic particles magneto-optical flux sensors are being investigated in order to enable direct visualization of stray flux.

Flux Method

Principle. The object becomes a part of a magnetic circuit. The flux in this circuit is generated by a permanent magnet or an electromagnet. The measured flux magnitude depends on the object's cross section and its magnetic permeability.

Probes. The flux in the circuit is measured directly by flux gates, magnetoresistors, Hall sensors or mechanical forces caused by the flux.

NDT. To detect and quantify corrosion damage that reduces the cross section of ferritic steel components magnetic yokes are guided over the surface. Assuming that the permeability of the material is constant the flux only depends on the cross section of the object. To locate the corrosion the flux magnitude can be mapped.

NDE. Under the assumption of constant permeability, the wall thickness or cross section of ferritic steel components such as sheets, pipes or cables may be assessed. Measurement of magnetic permeability becomes possible with components of sufficient thickness and lateral dimensions starting from a few square centimeters. The sensors pick up the degree of flux deflection caused by the ferromagnetic object. The same principle may be used for thickness measurements of nonferromagnetic walls. An additional ferromagnetic body (mostly a steel ball) placed on the backside deflects the magnetic field [3.73].

Tendency. For the assessment of more distant ferromagnetic objects high-sensitivity superconducting quantum interference devices (SQUIDs) are being used.

Residual-Field Method

Principle. In a first step the ferromagnetic object or a part of it is magnetized by a strong direct field as close as possible to its saturation. In a second step the residual field (Fig. 3.88, remanent induction, remanence) is measured, carrying information about the presence of the object, its microstructure, dimensions, and orientation.

Setup and Probes. The object can be magnetized by a yoke, a permanent magnet or an electromagnet. The magnetization can include the whole object or can be limited to a small area of a few square millimeters. The

residual field is measured by flux gates, magnetoresistors, Hall sensors, or SQUIDs.

NDT. This method is suitable for the detection and characterization of ferromagnetic particles in nonferromagnetic surroundings such as splinters of cutting tools in nonferromagnetic pieces [3.75].

NDE. The residual induction strongly depends on the microstructure of the ferromagnetic steels or cast iron due to its correlation with the mobility of magnetic domain walls. Evaluating the residual induction allows one to assess heat treatment, toughness, hardness, surface hardening (Fig. 3.89, [3.74]) or even carbon content. Calibration is the most important feature for the success of this method and should be accomplished according to appropriate guidelines. The method allows fast automatic inspection using conveyer movement of the objects through a magnetization tunnel and along a sensor station. Usually the object must be demagnetized afterwards.

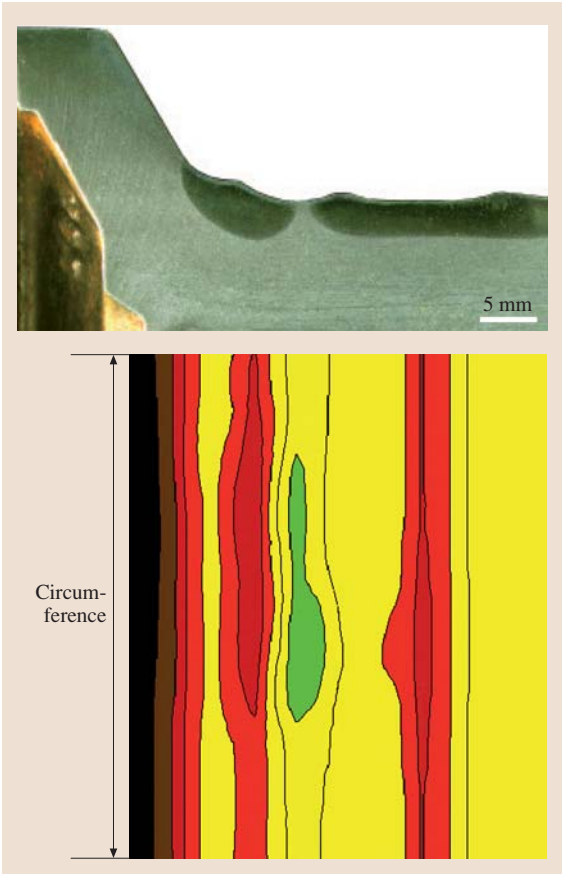


Fig. 3.89 Imperfectly hardened surface layer and the residual field distribution along the circumference (after [3.74])

Tendency. For reduction of the dimension's influence on the NDE results the coercive force may be evaluated. For this, after magnetization, the object is demagnetized by a contrary field and the strength of the demagnetizing field when the residual field vanishes is recorded.

Barkhausen Noise Analysis

Principle. The excitation of ferromagnetic material by a magnetic field that varies with time changes the spatial dimensions of the magnetic domains. The Bloch walls separating the domains from each other move discontinuously through the grain, emitting electromagnetic pulses. The superposition of these pulses produces a noise-like signal called magnetic Barkhausen noise [3.76]. The amplitude and rate of these pulses are discontinuously distributed over a complete magnetization cycle. Close to the coercivity they reach their maximum.

Setup and Probes. Excitation is accomplished by a magnetic yoke placed on the object. The driving coil is fed by an alternating current of a frequency ranging from a few tenths of a Hertz to a few hundred Hertz. Figure 3.90 displays a yoke and a sensing coil between the yoke limbs to receive the emitted pulses.

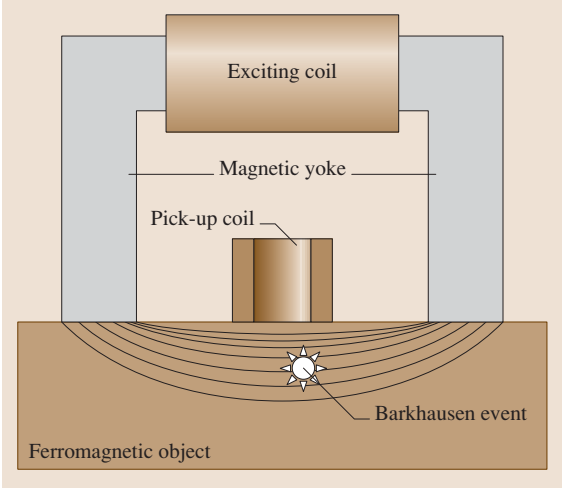


Fig. 3.90 Barkhausen noise excitation and measurement

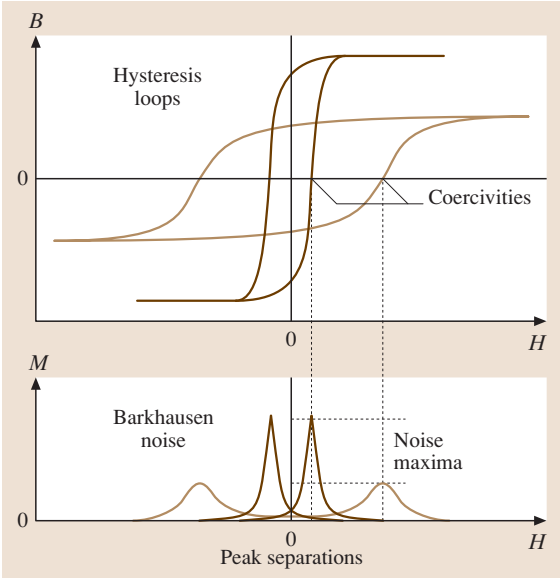


Fig. 3.91 Assessment of surface hardening depth of ferromagnetic steel (after [3.77])

NDE. The amplitude and pulse rate of Barkhausen noise depend on the Bloch wall mobility at different field strengths. This mobility is influenced by load and residual stress, grain and phase boundaries, grain orientation, and microstructural defects such as vacancies, dislocations, precipitates, segregations, and inclusions. The method is used to estimate hardness obtained by lattice defects, laser- and case-hardening depth, and stress state.

Figure 3.91 shows the principle of thickness assessment of surface-hardened layers in ferromagnetic steel.

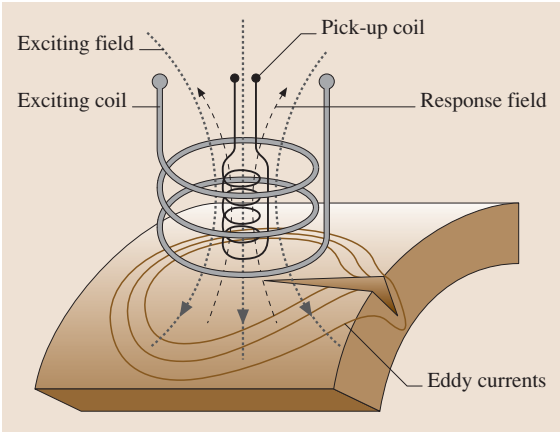


Fig. 3.92 An eddy-current surface probe detects cracks due to the deflection of eddy-current lines

The wide hysteresis loop describes the magnetic properties of the surface layer while the narrow loop results from the bulk material. During cyclic magnetization, different maxima of the Barkhausen noise amplitude may be observed [3.77]. At lower field strength the bulk material provides the first maximum followed by the maximum of the surface layer. The ratio between these maxima correlates with the hardened layer depth [3.78].

Tendency. Frequency analysis of the noise signal yields the source depth of the pulses. In combination with other micromagnetic parameters such as incremental permeability, local remanence or tangential field strength the method will find a wide field of applications.

3.5.5 Electromagnetic Methods

Electromagnetic methods not only rely on magnetic properties but also on the behavior of the material in an alternating electric field. For conductive materials such as metals even low-frequency electromagnetic fields induce an electric current. For nonconductive materials such as most ceramics and plastics, higher frequencies ($> 10\,\text{MHz}$) are necessary to generate a so-called dislocation current caused by various polarization mechanisms in the atoms or molecules.

Eddy-Current Method

Principle. An alternating magnetic field between 10 Hz and 10 MHz is applied to a conductive material. This field induces a circular voltage that drives a circular current with alternating direction almost parallel to the surface. As Fig. 3.92 shows, this so-called eddy current builds up its own magnetic field that counteracts the source. The sensor evaluates the resulting field, which contains information about the magnetic permeability, conductivity, and geometry parameters [3.73].

Probes. The probe consists of a transmitter generating the alternating magnetic field and a receiver to pick up the resultant magnetic field. The transmitter is commonly a coil; the receiver may be a coil or another magnetic field sensor such as a magnetoresistor, a flux gate, or even a SQUID for very low frequencies. Most simple sensors combine the transmitter and receiver in a single coil. For direct visualization of eddy currents in flat surfaces their magnetic field can be picked up by magneto-optic sensors such as garnet films.

NDT. Surface crack detection in all metals even below nonconducting coatings is the most common field of

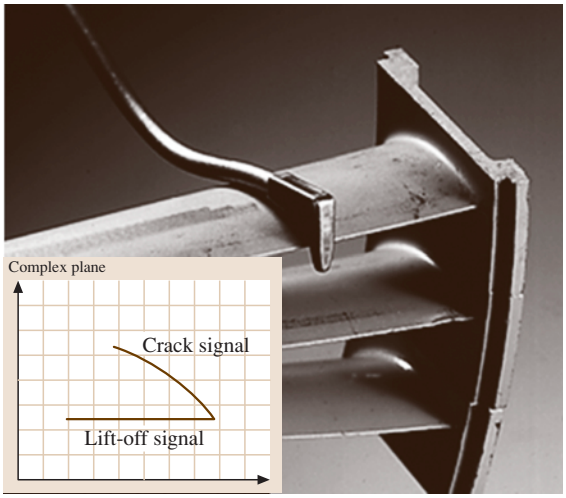


Fig. 3.93 Eddy-current inspection of the trailing edge of a turbine blade

application [3.79]. The detection and quantification of hidden defects such as pores, corrosion, and cracks is also possible in nonferromagnetic materials up to a few millimeters below the surface. Figure 3.93 shows the example of eddy-current turbine blade inspection.

NDE. The eddy-current method is best suited for conductivity measurement in nonferromagnetic materials and heat treatment characterization of pure metals and alloys. Material sorting can be accomplished as well as thickness assessment of nonconducting layers on conducting bulk or wall thickness assessment of nonferromagnetic sheets or pipes. In conducting composites such as carbon fiber reinforced plastic (CFRP) fiber orientation can be evaluated using the anisotropy of conductivity.

Tendency. Array sensors provide a fast and convenient opportunity to visualize eddy-current behavior. Highly sensitive and resolving sensors make smaller defects visible at greater material depth.

Microwave Method

Principle. Electromagnetic fields excited at frequencies from a few gigahertz to a few hundred gigahertz provide wavelengths in the centimeter and subcentimeter range, so called microwaves. These waves are reflected at the surface of metals but can penetrate many nonmetals, such as plastics, ceramics, and composites. various polarization mechanisms of the material components change the amplitude, phase, and polarization of the microwaves.

Setup and Probes. Mostly horn aerials are used to transmit and receive microwaves. Single- or double-sided aerials allow reflection and transmission measurements. For near-field applications an aperture may shape the field transmitted from the antenna. Directional couplers, phase shifters, and modulators complete the equipment [3.80].

NDT. Flaws in metals may be detected and characterized if they break through the surface. At very high frequencies wave propagation in open cracks may be used for crack depth estimation. In dielectric material internal flaws such as pores and delaminations may be detected due to the scattered energy.

NDE. Nonconductive materials can be inspected for material composition, structure, density, porosity, homogeneity, orientation of reinforcing fibers, state of cure, and moisture content. The reinforcing components in concrete of buildings may be visualized. For metals only thickness measurement of plates becomes possible using a double-sided reflection technique [3.81].

Tendency. Smaller electronic devices enable the integration of increasing numbers of components into one instrument, so that handling problems are decreasing. With increasing frequencies smaller defects will become detectable.

3.5.6 Thermography

Principle

Heat storage and transport capabilities depend on the heat capacity and thermal conductivity of the material as well as the local geometry of the object. Stimulated heat transport is used to evaluate the material for homogeneity, isotropy, and defects. For noncontact assessment the temperature distribution on the object is recorded using the thermally induced electromagnetic radiation of the object, which starts from wavelengths of about 10 μm at room temperature and can be visualized by using infrared sensors.

Setup and Probes

Figure 3.94 shows that a dynamic heat flow can be generated by periodically activated external (or internal) sources. External sources such as lamps, electric heaters, fans, and liquids heat the object from the front or back side. Internal sources may be stimulated by vibration or electric current. The temperature of the surface is recorded by infrared cameras based on scanning point, line or matrix sensors [3.82, 83].

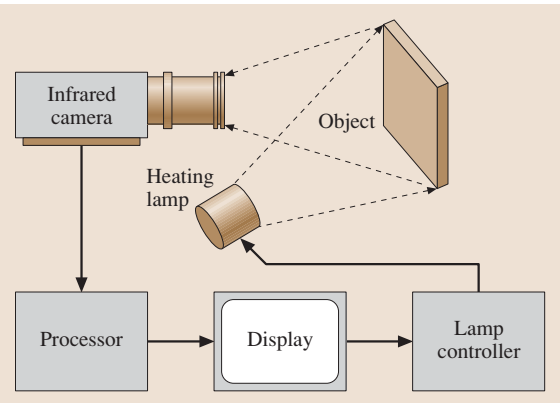


Fig. 3.94 An infrared camera picks up the temperature profile of the surface. The object can be heated by lamps (after [3.82])

NDT

Lock-in thermography is based on harmonic excitation of the heat source. The surface temperature of the object is analyzed according to its amplitude and phase. The phase signal is free of disturbances resulting from the emissivity of the surface. Another approach is the pulsed heating of the object and the time-dependent analysis of the surface temperature. Both methods are able to detect and characterize inner flaws like delaminations in composites, impact damage and debonding of joints. Figure 3.95 presents an application for turbine blades.

Defect selection may be achieved by activating cracks as heat sources. To do this, powerful acoustic waves are fed into the object. A local temperature increase due to face friction indicates the presence of a crack.

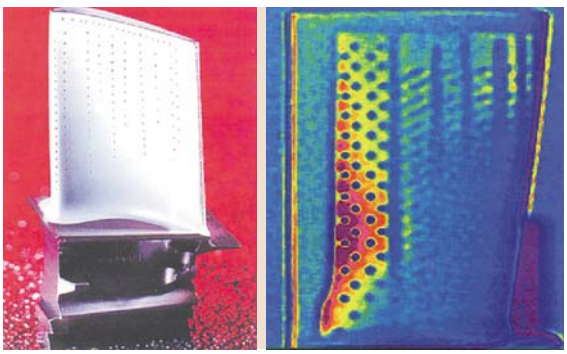


Fig. 3.95 Turbine blades with heat protection layer and cooling channels. Thermography highlights clogged channels

NDE

The thickness of surface layers on conducting and nonconducting bulk material may be addressed as well as the orientation of carbon fibers in CFRP.

Tendency

Current investigations are focused on defect-selective thermography, the assessment of the quality of adhesive joints, and composite materials. The increasing efficiency of infrared sensors is making infrared cameras increasingly affordable and convenient.

3.5.7 Optical Methods

This subsection summarizes methods based on visible light that is reflected from the object surface. Most attention has to be paid to the illumination and the visual abilities of the operators. Clear instructions and master pieces of what to look for are required.

Visual Methods

Principle. The object surface is cleaned and systematically searched for defined patterns corresponding to cracks, corrosion, microstructure or other features.

Probes. Many tasks are solved by the naked human eye. If necessary, lenses, microscopes, endoscopes, and appropriate recording instruments are used. The incident, intensity, and color of the illumination have to be optimized for the inspection task.

NDT. Without optical enlargement only large surface-breaking defects may be detected. For maintenance of engines, gear boxes, and other nearly closed hollow objects, endoscopes combining illumination, sensors, and sensor controllers are used [3.84]. Despite the distorted aspect ratio of the recorded picture it is possible to measure the dimensions of the visual pattern.

NDE. For the analysis of structural features the object has to be carefully prepared, including mechanical surface treatment such as grinding, polishing, and if necessary etching. The pattern can be interpreted under defined illumination and requires long experience.

Tendency. Tube-based endoscopes are being substituted by fiber and video endoscopes. Additionally, endoscopes may carry sensors for other NDI methods such as eddy-current probes. Some endoscopes allow the use of mechanical tools to treat any defects found.

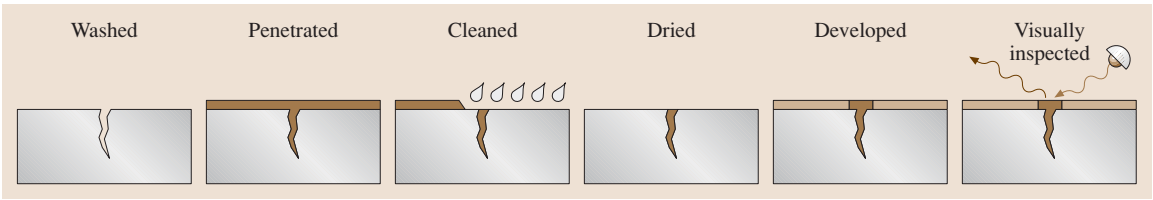


Fig. 3.96 For penetration inspection the object is first washed, then penetrated, cleaned, dried, developed, and visually inspected

Penetration Methods

Principle. Figure 3.96 illustrates the operation of this method. The cleaned surface of the object is coated with a penetrant in which a visible or fluorescent dye is dissolved or suspended. The penetrant is pulled into surface cracks by capillary action. After cleaning the surface of excess penetrant a developer is sprayed or dusted over the object, partially lifting the penetrant out of the crack. Under defined illumination the penetrant provides an enlarged crack pattern with high contrast [3.85]. A roller with surface cracks is shown in Fig. 3.97.

Equipment and Inspection Agents. For manual inspection spray cans with the penetrant and the developer are used. In modern inspection lines all the objects are washed, penetrated, cleaned, and developed automatically at defined agent temperatures and action times in immersion tanks. Visual inspection for defect indications is performed by human operators under either visible or ultraviolet illumination.

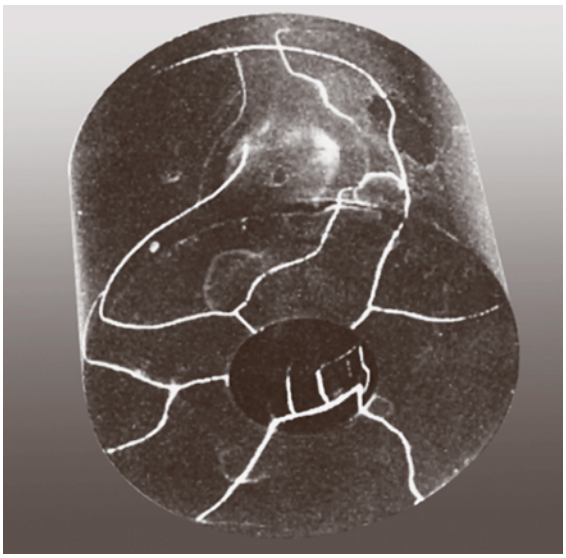


Fig. 3.97 Result of penetrant inspection for cracks in a roller

NDT. The crack indication varies with developing time. Reference master pieces with known defects and exact instructions enhance the reliability of this method. No information about defect depth can be obtained. Post-emulsifiable penetrants keep the viscosity at a low level over time.

Tendency. To date the interaction of the operator is needed to distinguish between real defects and pseudo-indications. Much effort is focused on substituting the inspector with an automatic vision system.

Speckle Interferometry

Principle. Speckle interferometry uses the interference phenomena of laser light. Figure 3.98 shows that the object is entirely illuminated with laser light, producing a speckle pattern on the object surface. This pattern is superimposed by reference light and the resultant image is recorded by a video camera. For NDI this method is used to detect and quantify surface dislocation at loading.

Setup and Probes. Laser illumination is performed by a defocused laser so that the entire surface is illuminated at once. No scanning is necessary. To smooth the illumination the laser light may reach the surface via several

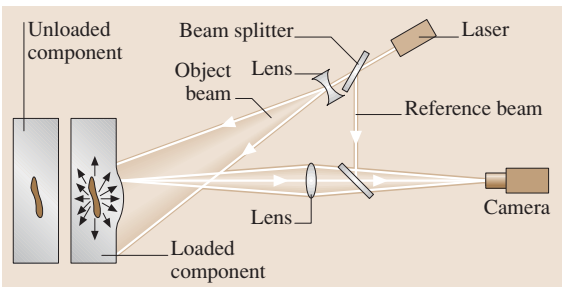


Fig. 3.98 Electronic speckle interferometer (after [3.82])

Fig. 3.99 Speckle fringes caused by discontinuities of off-plane surface displacements

paths. A beam splitter directs a small portion of the light as a reference to the camera for interference with the reflected light from the object. The camera records the images and passes them to data storage.

NDT. The speckle pattern of the object is recorded twice. The correlation of the images of the unloaded and loaded object highlights discontinuities of displacements penetrating the surface and allows local debonding, delaminations, and cavities to be recognized as fringes. An example is shown in Fig. 3.99.

Tendency. The differential speckle interferometry method known as shearography highlights differences in surface dislocation of points at a certain distance from each other. The method is less sensitive to disturbances and best suited for in-field inspection [3.86].

3.5.8 Radiation Methods

High-energy radiation is able to penetrate solid bodies and to interact with their atoms. The transmitted intensity depends on the atomic number, the density of the material, and its thickness. The object is illuminated entirely and imaging can be performed by films or electronic matrix imagers [3.87, 88].

X-Ray Method

Principle. An X-ray tube generates radiation with energy up to a few hundred keV. The radiation is directed to the object positioned a certain distance from the tube. A radiographic film or conversion screen close behind the object records the transmitted radiation intensity as a grey scale image.

Equipment. An X-ray source is shown in Fig. 3.100. It consists of an evacuated tube in which a cathode

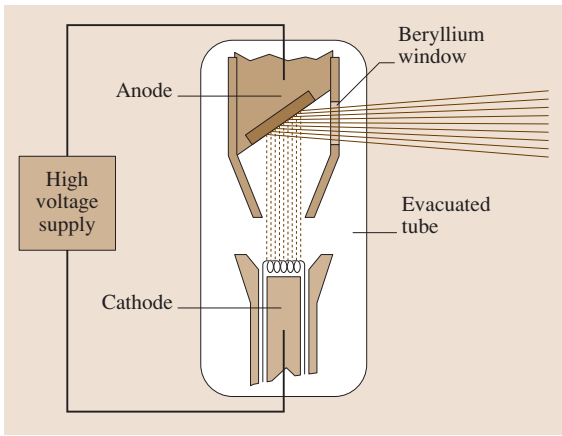


Fig. 3.100 Scheme of an X-ray tube (for explanation see text)

emits electrons that are accelerated towards the anode. The electrons strike the anode and emit bremsstrahlung, i. e., X-rays with a continuous range of energies. This energy can by controlled by the voltage between the cathode and anode. The radiation leaves the tube via a beryllium window and radiates the object. Increasingly, conversion screens and storage foils are uses for imaging.

NDT. Defect detection is based on alteration of the X-ray attenuation by the defect. Depending on the defect material, this attenuation may be smaller or greater than in its absence, so that either increased or decreased X-ray intensity can be detected. While the detection of volume defects such as pores starting from a defined extension is very reliable crack detection requires their correct ori-

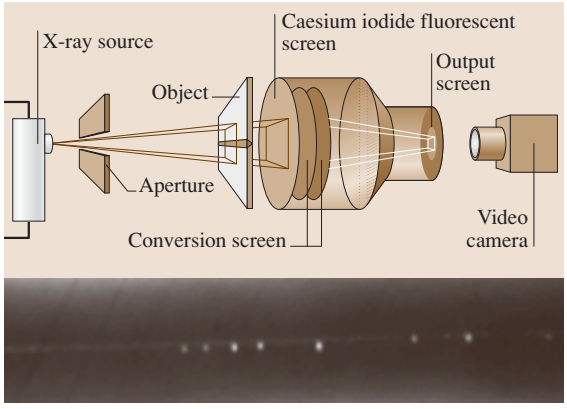


Fig. 3.101 Radioscopic equipment and X-ray image of a weld with pores

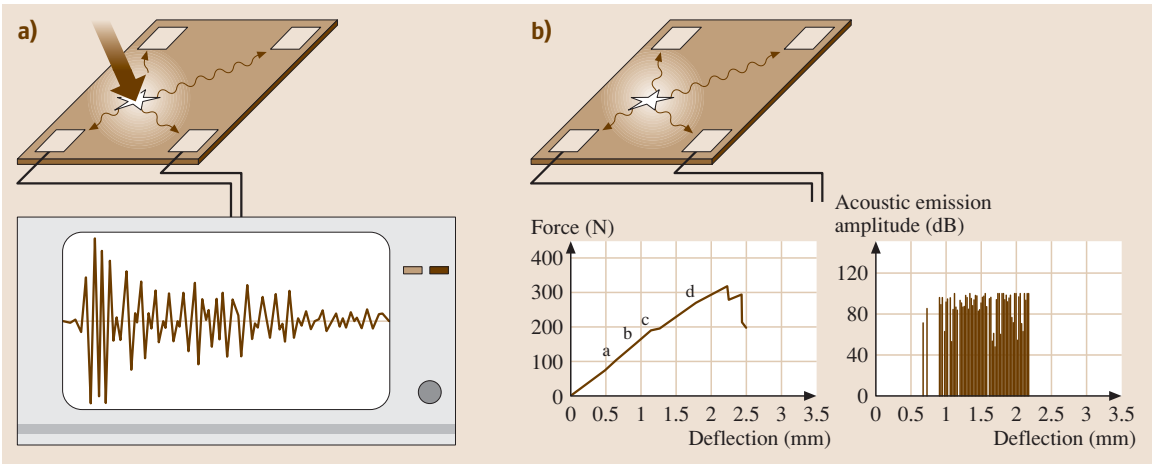


Fig. 3.102a,b Damage processes can be detected by their acoustic emission. Impact damage (**a**) or crack growth (**b**) can be detecting and localization (after [3.89])

entation in the X-ray beam. Figure 3.101 presents the equipment and an X-ray image of a weld. For defect detection in steel X-ray energy of up to 500 keV is required to penetrate walls of 100 mm thickness.

NDE. X-rays are used to gauge the wall thickness of pipes and sheets. With a multi-energy technique it is possible to detect the atomic order of the material, which is used for material identification.

Tendency. To reduce the blur of X-ray images microfocus tubes are used. For cross-sectional imaging computer tomographs are used, turning the object in the X-ray beam. Tubes with turning anodes are needed to increase the X-ray intensity by increased electron current in the tube.

Gamma-Ray Method

Principle. The radioactive decay of some elements produces high-energy gamma rays that are able to penetrate metals to a thickness of a few centimeters. Small pellets measuring a few millimeters are activated in a nuclear reactor and then stored in highly damping containers. These continuously radiating pieces are called sources. The decrease of their activity with time is described by their half-life constant and depends strongly on the source material.

Equipment. For exploitation the source is loaded into a mobile source holder made from a dense material such as tungsten, uranium or lead. Via remote control this holder is opened and the source is moved out to radiate the object through a beam collimator. The transmitted radiation is recorded by using a radiographic film.

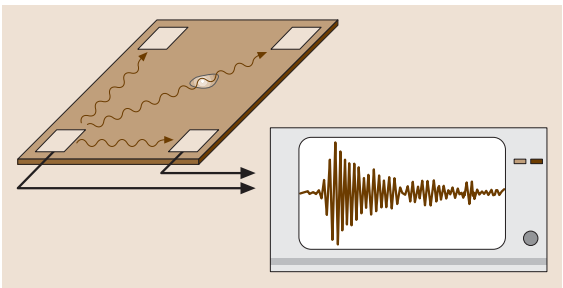


Fig. 3.103 Acousto-ultrasonic measurements (*left*) reveal defects between transmitter and receiver

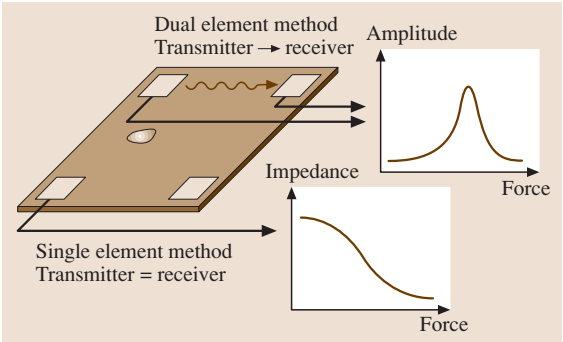


Fig. 3.104 Impedance spectroscopy allows the detection of defects on or close to the piezoelectric element (after [3.90])

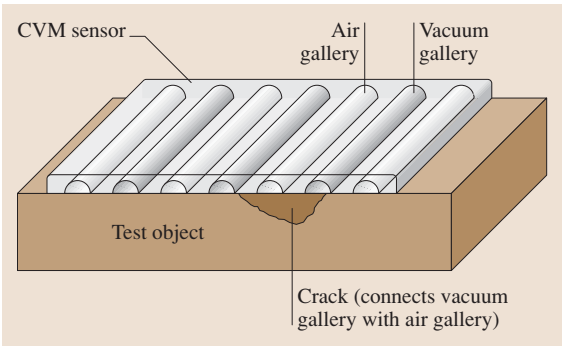


Fig. 3.105 Comparative vacuum monitoring (CVM) detects defects due to varying differential pressure (after [3.91])

NDT. For very thick steel components cobalt-60, which radiates with an energy of more than 1 MeV, is used. For steel thickness of less than 50 mm softer sources such as iridium-192 are sufficient. The safety requirements for X- and gamma-ray exploitation are very restrictive due to their harmful interaction with biologic tissues.

Tendency. Selenium-75 is best suited for steel walls up to 30 mm and has a significantly longer half-life. This source is able to replace X-ray equipment.

3.5.9 Health Monitoring

(SHM) *Structural health monitoring* refers to nondestructive inspection methods that rely on integrated sensors in the inspected structure itself. The sensor signals may be monitored online at the loaded structure or recorded for offline analysis. Various such NDI techniques are being investigated for applications in aircrafts [3.92, 93], buildings [3.94], and power stations. Active sensors can transmit and receive signals while passive sensors receive signals generated by the damage process or damage growth.

Acoustic emission (Sect. 3.4.2) can be recorded by embedded or attached piezoelectric sensors. As shown in Fig. 3.102 the source of an emission can be an impact, the growth of cracks, fiber or matrix breakage, delamination, and other damaging processes. To localize these sources signals from different sensors are correlated, yielding differences in time of flight for use in triangulation algorithms.

Acousto-ultrasonic interrogation is a single-sided nondestructive inspection technique employing separated sending and receiving transducers (Fig. 3.103). The method is used for assessing the microstructural condition and distributed damage state of the material between the transducers [3.70].

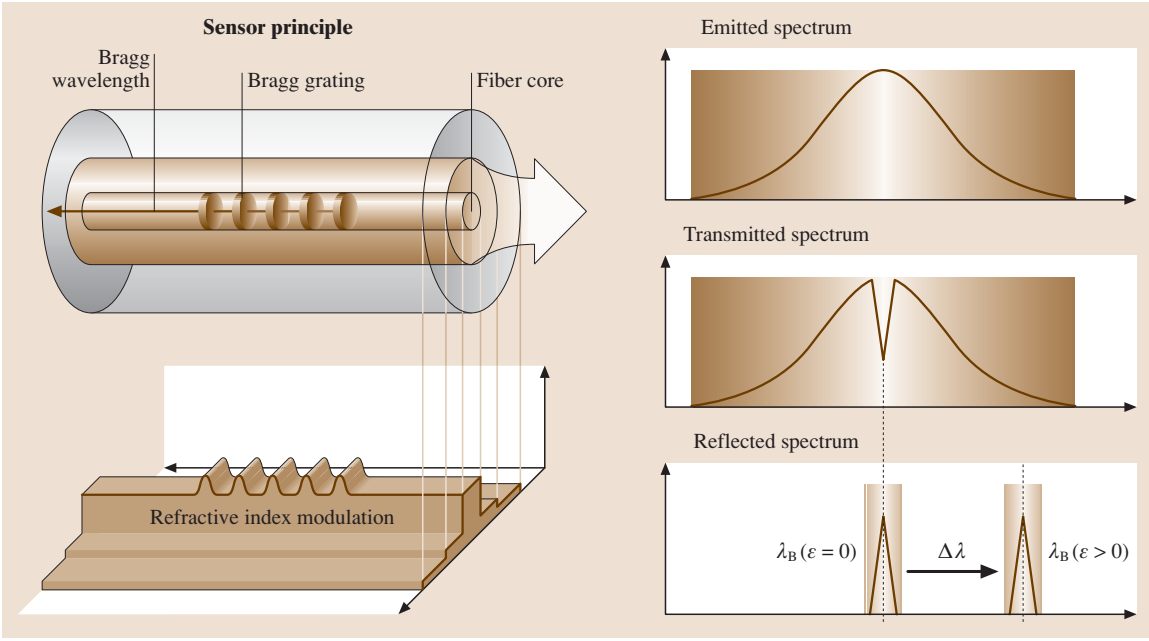


Fig. 3.106 Fiber Bragg grating (after [3.91])

Impedance spectroscopy uses either a single piezoelectric element or a transmitter–receiver combination (Fig. 3.104). The excitation oscillates in a predefined frequency band and the measurement is either the impedance or the complex voltage at the receiver. The frequency-dependent behavior of the measurement indicates defects on or close to the piezoelectric element [3.90]. Both, acousto-ultrasonic and impedance spectroscopy can be used to inspect polymer matrix composites, metal matrix composites, ceramic matrix composites, and even monolithic metallic materials.

Eddy-current foil sensors are an alternative technology to the classical eddy-current technique (Sect. 3.5.5) for the detection of surface or hidden cracks. In this method, a copper winding is printed onto a plastic substrate, just like an electronic track. Due to their thin geometry, they can be mounted onto interfaces between structural parts, around bolts, in corners, and hardly accessible regions. Periodic reading of these coils can provide information on structural health.

Comparative vacuum monitoring offers an effective method for in situ real-time monitoring of crack initiation and/or propagation. This method measures the differential pressure between fine galleries containing a low vacuum alternating with galleries at atmosphere in a simple manifold (Fig. 3.105). Comparative vacuum monitoring enables the monitoring of the external surfaces of materials for crack initiation, propagation, and corrosion. The galleries can also be embedded between components or within material compounds such as composite fiber.

Fiber Bragg gratings measure either the tensile or compressive strain applied along the grating length of an optical fiber (Fig. 3.106). The grating consists of a periodic variation of the index of refraction and provides a linear relationship between the change in wavelength of the reflected light and the strain in the fiber caused by externally applied loads or thermal expansion. To operate multiple sensors along a single optical fiber, the various Bragg gratings should have different Bragg wavelengths in order to differentiate between them.

3.6 Corrosion

3.6.1 Background

In general, corrosion is understood to refer to material degradation through reaction with its environment. This has led to a common tendency to assess it in terms of the corrosion products which are formed, i. e., concentrating on the phenomenon rather than its cause. Recent developments in observing and measuring corrosion are increasingly changing this picture. As a result, it is necessary to give up commonly held assumptions in order to understand the nature of corrosion. Among other things, the order of standard potentials of the elements has been overemphasized for some time in terms of its relevance.

In contrast to the other topics described in this Chapter, it is hardly possible to describe the corrosion behavior of technical equipment and structural components by means of formulae, tables or guidelines. The reason for this is that their corrosion resistance, and thus corrosion itself, is not just a property of the material, but rather of the system as a whole. The actual corrosion behavior is dependent in equal measure on the metal (as a technical material, taking into account all its properties), the environment (i. e., the concentration, temperature, flow rate, etc. of the corrosive medium), and the equipment design. In this context, design has to be understood in a broader sense to encompass everything from microscopically small surface roughness, methods of joining parts together, combinations of materials (including crevices resulting from the design) right through to the equipment construction as a whole. As a result, a large number of influencing factors are involved and the possible variations become difficult to comprehend. Thus corrosion behavior always has to be assessed in terms of the character of the complete system, and a so-called *corrosion atlas* is of little help. Even if the appearance of material damage is similar in more than one case, this does not mean that the causes are the same.

In practice, the cumulative experience gained from failures, one's own technical knowledge, and the corrosion data to be found in the literature always possess validity only over a narrow range of situations. Small deviations in particular parameters (locally reduced concentration of oxygen with stainless steels, shifts in the pH value with aluminum, attainment of a critical temperature level, etc.) can have dramatic consequences. A number of physical factors, such as

Table 3.8 Energy required to produce metals from the compound state and the standard potential E_0 at 25 °C within the order of potentials of individual elements (SHE = standard hydrogen electrode), see also [3.95]

Metal	Metal oxide	Energy required for production (kJ/kg)	(kJ/mol)	Standard potential (mV) (SHE, 25 °C)
Al	Al_2O_3	29 200	788	−1660
Cr	Cr_2O_3	10 260	534	−740
Fe	Fe_2O_3	6600	367	−440
Ni	NiO	3650	213	−250
Pb	PbO	920	191	−130
Cu	Cu_2O	1180	75	+340
Ag	Ag_2O	60	6	+790
Au	Au_2O_3	−180	−37	+1500

mechanical stresses or the uptake of solvents leading to swelling of plastics, also have a strong influence on corrosion behavior. This virtually unlimited spectrum of influencing factors and conditions cannot be accommodated within rigid guidelines. Instead, it is important to become acquainted with the nature of corrosion itself (and with its apparent contradictions) in order to be in a position to assess the risk in a concrete situation, or to clarify specific aspects in cooperation with experts, sometimes by carrying out appropriate experiments.

Corrosion can be divided into two main types:

1. Electrochemical corrosion (the atmospheric corrosion of steels, often equated with rusting, is an important example here)
2. Chemical corrosion (high-temperature corrosion, leading to scale formation on steels, is a key area here, but the corrosion of glass, ceramics, and concrete is also primarily chemical in nature)

3.6.2 Electrochemical Corrosion

Fundamentals

In order to understand corrosion, it is vital first to consider its ultimate cause, i. e., the driving force. Most common metals are produced under the expenditure of large amounts of energy from their compounds, mostly oxides; for example, 6600 kJ/kg are required to produce iron from Fe_2O_3 and as much as 29 200 kJ/kg to produce aluminum from Al_2O_3. Further examples are given in Table 3.8. The durability of metals is thus limited by nature, since the material always attempts to attain a condition of lower energy. In general, the conversion back to this state occurs more quickly, and the tendency for this to happen is higher, the further away the metal is from the energetically stable condition. Human efforts to prevent this are limited to influencing the kinetics of the reconversion and delaying the attainment of the thermodynamically stable, nonmetallic state. This can be achieved over an appropriate period of time by means of various measures, the use of coatings being one such example.

If a metallic surface comes into contact with water, the process of metal dissolution begins spontaneously. During this process, the metal goes into solution as an ion (Me^{z+}) and, depending upon its valence (z), one or more electrons (ze) are set free and remain within the metal. The release of electrons is also known as oxidation. Note, however, that oxidation is not necessarily associated with oxide formation. The originally neutral metal becomes negatively charged via the electrons left behind during this process and thus the dissolution can be described electrically by means of Faraday's law

$$\Delta m = \frac{MIt}{zF}(g)\,. \tag{3.80}$$

In (3.80), Δm is the loss of mass, M is the molarity, I is the flow of electrons (current amplitude) as a result of metal dissolution, t is time, and F is Faraday's constant. If the electrons are not consumed, charge separation rapidly leads to an increase in electrostatic forces, which then prevents further metal dissolution. Thus a so-called dynamic equilibrium is attained, in which the same number of metal ions undergo dissolution as are returned to the metallic state

$$\mathrm{Me} \leftrightarrow \mathrm{Me}^{z+} + z\mathrm{e}^{-}\,. \tag{3.81}$$

In analogy to a plate condenser, the charge in the metal (free electrons) is opposed by an equivalent level of positive charge within the electrolyte (Fig. 3.107). This electrolytic double layer is the location of the potential difference between the metal and the electrolyte, i. e., the electrode potential E. This potential can

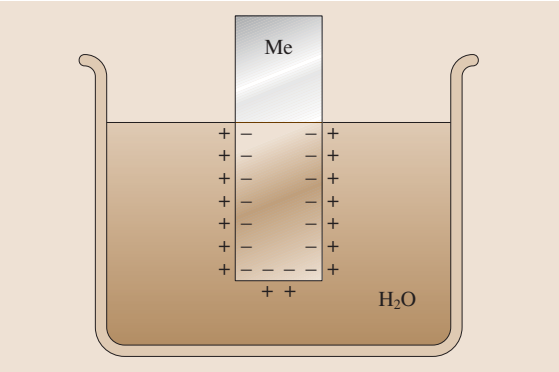


Fig. 3.107 Formation of an electrolytic double layer at the phase boundary metal/water; the more metal ions enter solution, the more negative the metal becomes

only be determined indirectly with the aid of a reference electrode (e.g., the standard hydrogen electrode or a calomel electrode). The size of the electrode potential (the charge separation) depends upon the metal, the valence, the temperature, and the natural logarithm of the concentration of metal ions already present in solution.

This behavior was summarized in an important equation, named after its discoverer, the German physicist and physical chemist Nernst (1864–1941)

$$E = E_0 + \frac{RT}{zF} \ln c_{Me^{z+}}(V) \,. \qquad (3.82)$$

In this equation, E_0 is the so-called standard potential, and c_{Me}^{z+} is the concentration of ions of the relevant metal in solution. If a metal is inserted into an aqueous solution where the concentration of its own ions amounts to 1 mole per liter, the right-hand term in the equation becomes zero (since $\ln 1 = 0$) and $E = E_0$. The standard potentials E_0 can be found in the table of standard electrode potentials of the elements (Table 3.8). These demonstrate very clearly the correlation between the energy expended and the tendency to return to the energetically lower state. The table of standard electrode potentials is *unsuitable*, however, as a basis for assessing practical corrosion behavior, since entirely different parameters (medium, alloying elements, film formation, area ratios, etc.) play the dominant role here. Up to this point, only a homogeneous electrode has been considered. In practice, however, metals represent technical materials which are anything but homogeneous. The presence of impurities and/or alloying elements (either in solution or as precipitates), the existence of different heat treatment states, levels of deformation, different protective or adsorbed layers, crystallographic anisotropy, and various lattice defects all lead to the creation of locations with different (usually higher) energy. The tendency of the metal to return to an energetically lower state is thus particularly high at such locations.

As discussed above, electrostatic forces between the free electrons and the metal ions in solution prevent further dissolution of the metal. This only becomes possible through a process which consumes electrons, whereby only four reactions need to be considered. Unfortunately these are generally present and lead to various different types of corrosion:

- Discharge of hydrogen ions (corrosion in acids)

$$2H^+ + 2e^- \rightarrow H_2$$
$$\text{(more exactly:} \quad 2H_3O^+ + 2e^- \rightarrow 2H_2O + H_2) \qquad (3.83)$$

- Reduction of oxygen dissolved in the water (atmospheric corrosion leading to the formation of rust via subsequent reactions)

$$O_2 + 2H_2O + 4e^- \rightarrow 4OH^- \qquad (3.84)$$

- Deposition of more noble metallic ions (corrosion through use of mixed metals)

$$\text{e.g.,} \quad Cu^{2+} + 2e^- \rightarrow Cu \qquad (3.85)$$

- Applied current (e.g., corrosion through stray currents in the Earth near tram lines)

The process which consumes electrons does not have to occur at the location of metal dissolution, but can also occur in an entirely separate place which is more favorable for electron transfer. Thus, the popular description *consumed by acid* is very misleading, since

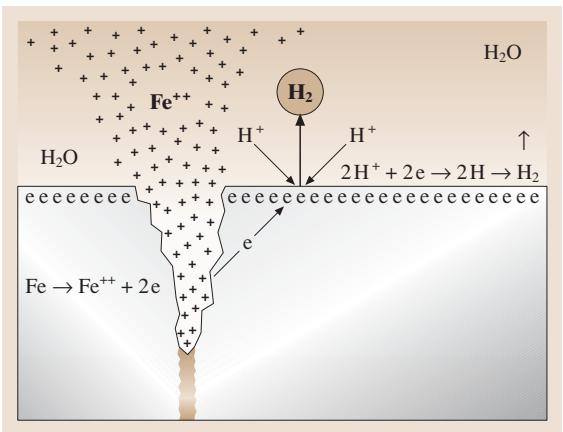


Fig. 3.108 Corrosion at a grain boundary (schematic)

the metal atoms can continue to leave their lattice as ions *only through the take-up of electrons* and this will take place at locations of higher energy, such as grain boundaries, hardened regions, etc.. The concept of acid *eating into metal* is, in any case, false since the positive hydrogen ions would have to flow into the regions of positively charged metal ions, a process that would be made impossible by the forces of electrostatic repulsion. Figure 3.108 shows schematically the dissolution of metal at a grain boundary. It is intended to make clear where metal dissolution occurs and demonstrate that the effect of acids (via an electron-consuming process) actually takes place elsewhere, namely at the grain surfaces.

The misfit in the grain-boundary region is larger, the higher the orientation differences between adjacent grains. This is why metal ions can easily leave such locations, provided that electron consumption is possible. It also explains the fact that grain boundaries appear to be of different width following metallographic etching. The spatial separation between the location for dissolution (anode) and for electron consumption (cathode) can become relatively large if multiphase materials are involved, or particularly if components made of different metals are connected together so that electron conduction is possible (see bimetallic corrosion). This effect is used to great advantage in batteries. Inhomogeneities can also exist in the medium (the environment surrounding the metal) in an analogous way to the inhomogeneities on the metal surface (which transform uniform into selective dissolution). Such differences can occur, e.g., through processes leading to a reduction (consumption of hydrogen ions) or increase (alkalization through the formation of OH^-ions) in the concentration of species, diffusion processes (transport to and from the surface, as well as in the bulk), blockage of charge transfer via adsorbed layers, and secondary corrosion products (rust).

Figure 3.109 shows the influence of the thickness of a moisture film on atmospheric corrosion. If no water is present (moisture approaches zero), no electron-consuming process can take place in accordance with (3.84). As the amount of water increases, the electron consumption becomes more rapid and the metal can undergo dissolution more easily. After a certain film thickness (100 µm) is reached, however, the rate of corrosion again decreases, since the oxygen from the atmosphere now has a longer diffusion path and consequently cannot consume so many electrons. In practice, this means that condensation can be more effective in causing corrosion than rain, or that a motorcycle which is stored during the winter under a so-called *protective*

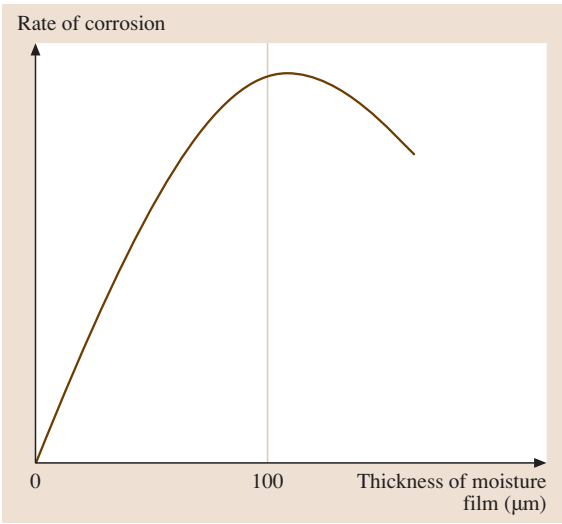


Fig. 3.109 Influence of the thickness of a moisture film on the rate of atmospheric corrosion (after [3.96])

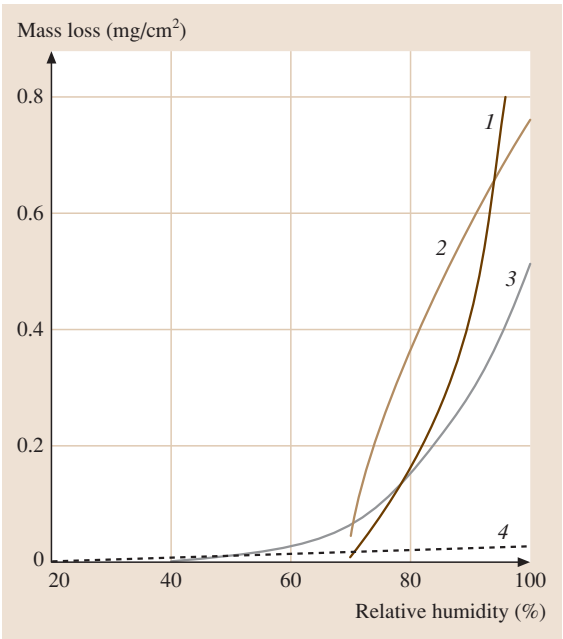


Fig. 3.110 Influence of air humidity on the corrosion of iron at constant temperature: 1 – air with 0.01 vol. % sulfur dioxide, and solid particles; 2 – air with 0.01 vol. % sulfur dioxide; 3 – air with solid particles; 4 – pure air (after [3.97])

cover can be exposed to ideal conditions for corrosion as a result of poor air circulation.

Atmospheric Corrosion, Rusting

Equation (3.84) indicates that atmospheric corrosion requires oxygen and water for the oxidation (electron consumption) of iron to occur. Oxygen is available in sufficient quantities from the atmosphere. With increasing temperature, the air can take up more water. This leads to a decrease in humidity for constant overall wetness (g/m^3). Conversely, the relative humidity increases (for the same wetness level) with decreasing humidity. The temperature at which the relative humidity is 100%, i. e., when the air is saturated with water, is called the dew point. Below this temperature, condensation occurs and leads to the formation of liquid water. Water can also form, however, at considerably lower relative humidity on air pollutants (dust particles) or on surfaces, since these both function as sites for the initiation of condensation.

Figure 3.110 shows the influence of air humidity on the corrosion of iron at constant temperature. It can be seen that the mass loss through corrosion increases dramatically at humidities above 70%. This relative humidity level is therefore referred to as the critical air humidity. If the humidity is lower, no significant corrosion occurs in practice. In very pure air, no real corrosion takes place even at a relative humidity of 100%. However, other factors such as air impurities (particularly sulfur dioxide) can lead to increased corrosion. Sulfur dioxide arises in large amounts from the combustion of organic fuels. Hygroscopic dust particles, such as corrosion products or airborne impurities (e.g., soot particles, salts), also favor condensation and lead to the formation of electrolyte films that permit corrosion to occur.

The formation of rust involves a secondary reaction following corrosion under atmospheric conditions according to (3.84). After iron ions have gone into solution, i. e., after corrosion has already occurred, these can react with hydroxyl ions (OH^-) and form $Fe(OH)_2$

$$Fe^{2+} + 2OH^- \rightarrow Fe(OH)_2 \,. \qquad (3.86)$$

Figure 3.111 shows the corrosion of iron under a water droplet. The diffusion path for oxygen is shortest at the edge of the droplet and this is where the electron-consuming processes take place with the formation of OH^- ions. The corrosion itself occurs at the center of the droplet, where iron atoms leave the metallic lattice and go into solution as ions. The electrons which are thereby set free

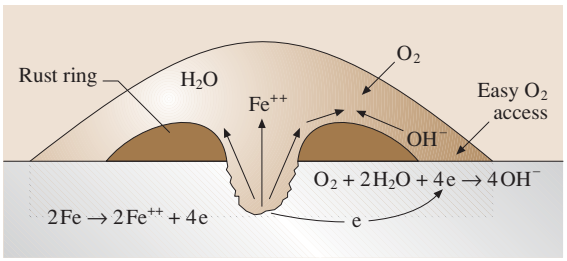


Fig. 3.111 Corrosion under a water droplet (after [3.98])

are consumed by reduction of oxygen. The iron and hydroxyl ions diffuse towards each other and initially form iron(II) hydroxide (3.86) which – in the presence of oxygen – is subsequently oxidized to γ-FeOOH (lepidocrocite) as the first crystalline product, according to

$$2Fe(OH)_2 + \frac{1}{2}O_2 \rightarrow 2FeOOH + H_2O \,. \qquad (3.87)$$

As the process continues over time, a number of further chemical reactions also occur and transform the initial rust into a mixture of different rust minerals. Depending upon its exact chemical composition, the volume of the rust is six to eight times larger than the missing (corroded) amount of iron. Lepidocrocite, also known as esmeraldite, is an unstable modification of the hydrohaematite FeOOH, which is contained primarily in the rust on low-alloy steels. It exhibits a red to brown color. The complex influence of climatic factors transforms the metastable form of lepidocrocite (γ-FeOOH) in part into the much more common rust mineral goethite (α-FeOOH). The name originates with Johann Wolfgang von Goethe, who first described this naturally occurring mineral. It is a further species of the metastable hydroxide FeOOH and is the primary reaction product on carbon steels. It exhibits a light yellow to blackish brown color. Direct transformation is hard to conceive, since lepidocrocite possesses the most dense packing of oxygen atoms in a cubic structure, while goethite has the most dense packing of oxygen atoms in a hexagonal structure. Instead, it has to be assumed that the transformation proceeds via the dissolution of lepidocrocite, followed by precipitation from a solution containing Fe(III). This assumption is supported by the 30 000 times better solubility of γ-FeOOH in comparison to α-FeOOH. An amorphous Fe(III) hydroxide forms as an intermediate product and is transformed into α-FeOOH by ageing. Higher temperatures accelerate this process. The composition of the corrosion products is also dependent upon the access of oxygen.

Table 3.9 Relative amounts of the crystalline phases as a function of the climatic conditions (selection) [3.99]

Climate	Rust composition		
	Lepidocrocite	Goethite	Magnetite
Industrial atmosphere	5	3	1
Marine atmosphere	0	2	3
Forest atmosphere	3	7	1

Magnetite (Fe_3O_4) is formed at those locations where component design limits the access of oxygen (corners and crevices). This so-called *dense rust* can cause plastic deformation or failure of component fasteners as a result of the increase in volume. It has a blackish brown to black color. The access of oxygen is also limited under thick layers of water.

A system of reversible reactions is formed [3.100]

$$Fe^{2+} + 2FeOOH \underset{O_2\ \text{deficiency}}{\overset{O_2\ \text{entry}}{\rightleftarrows}} FeO_4 + 2H^+ . \tag{3.88}$$

Depending upon the amount of water and the access of oxygen, the following phases can also be formed directly, provided prior oxidation of Fe^{2+} to Fe^{3+} has taken place

$$Fe^{3+} + 3OH^- \rightarrow FeOOH + H_2O , \tag{3.89}$$

$$2Fe^{3+} + 6OH^- \rightarrow Fe_2O_3 + 3H_2O . \tag{3.90}$$

Haematite (Fe_2O_3) forms under limited oxygen supply, either via reoxidation of Fe_3O_4 to Fe_2O_3, or by prior oxidation of Fe^{2+} to Fe^{3+} and subsequent reaction with hydroxyl ions. It exhibits a light brown color and has a cubic structure.

The relative amounts of the crystalline phases arising in the course of weathering can be determined by X-ray analysis, as shown in Table 3.9. These results are strongly dependent upon climatic conditions.

The factors already discussed demonstrate that the corrosion products on unalloyed steel are highly heterogeneous as a result of the variety of compounds which are formed, their ability to undergo transformation, and their different crystal structures. As a result they do not form adherent, protective layers on the steel.

Passivity

After very rapid, initial corrosion (e.g., as a consequence of plentiful oxygen supply in the water and the strong reactivity of the metal itself), a metallic surface can become spontaneously covered with an oxidic layer, a so-called passive film. Among others, the metals aluminum, titanium, zirconium, zinc, chromium, tantalum, cobalt, and nickel all fall into this category. Figure 3.112 illustrates what happens when these metals are exposed to the atmosphere, rather than under the special conditions of standard potential, absence of oxygen, and a one-molar solution of their own ions. Although thermodynamic properties are still valid, they disappear into the background and kinetics dominate the corrosion behavior as a result of the formation of passive films (which are electronically semiconducting or isolating). The only thing that then matters is what can still dissolve through the film and how the electron-consuming processes can occur. This characteristic transforms these nonnoble metals into the most important technical alloys.

For iron under atmospheric conditions, the reaction does not occur quickly enough: as a result, the dissolution process, rather than film formation, dominates and leads to the formation of undesired rust via secondary reactions. However, in alkaline media (e.g., in concrete where the pH value is > 12) or in strongly oxidizing acids (e.g., nitric acid), passivity can also occur spontaneously with unalloyed and low-alloy steels. In order to protect iron-based materials outside these special cases, they are alloyed with chromium to attain the desired state of passivity. This effect was discovered at the beginning of the 20th century by Maurer and Strauß while studying an experimental heat of steel (V2A) and was patented by the Krupp company in 1912.

Chromium is lower in the list of standard potentials than iron. Thus its addition as an alloying element actually makes the metal less not more *noble* in a thermodynamic sense. However, the passivity of the resulting alloy leads to it being referred to as

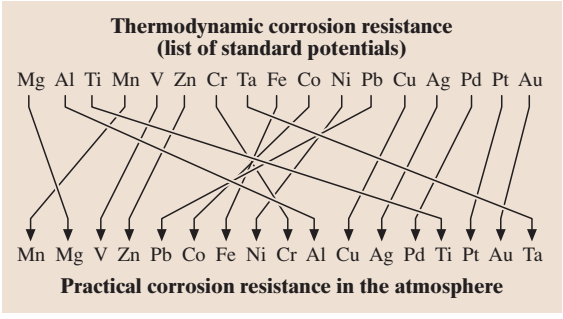


Fig. 3.112 The *nobility* of metals, thermodynamics versus kinetics (schematic)

a stainless steel. In German, such alloys are sometimes called *noble steels*, which is confusing with regard to corrosion and originates, in fact, from the way in which their phosphorus and sulfur contents were strongly reduced and low levels of nonmetallic inclusions were attained in the steel. Chromium's tendency to react very rapidly with the environment and form a subsequently protective film of corrosion product is transferred to iron–chromium or iron–chromium–nickel alloys once a concentration of approximately 12% chromium is attained in the material.

Passive films form naturally and their nature is thus very different from the properties of coatings or plated layers on steel. With stainless steels, the passive layer is remarkably thin: of the order of 10 nm (some 50 atom layers) and even less (five atom layers) with pure chromium. Such layers cannot be detected by conventional means, since their thickness is much smaller than the wavelength of visible light. The film is subject to very strong mechanical stresses and exhibits an extremely high potential gradient of 1 MV/cm. It is nonuniform in nature, both in terms of chemical composition and structure: adjacent to the metal, the film is amorphous, but it becomes increasingly crystalline towards the interface with the medium. Passive films are capable of repairing themselves after being damaged mechanically, which is of great practical importance and distinguishes them significantly from organic coatings, since the latter no longer provide an effective barrier once damaged.

In recent decades, research into the phenomenon of passive films has made astonishing progress. It is now known that the film does not correspond to an unchanging layer, but is part of a dynamic system. The relevant specialist literature now talks about the passive film *living*, whereby both birth and death events take place [3.102]. At any moment in time, activation and repassivation processes occur on a submicroscopic scale and statistically distributed over the surface. Under certain circumstances, these can be measured as small impulses of potential and current. For some years now, it has been recognized that these impulses, which are known as electrochemical noise, are dependent upon the nature of the metal and its actual state, the temperature, the pH value, and the type and concentration of ions dissolved in the medium. They provide an important source of information on corrosion behavior and electrochemical noise exists even when it is not being measured. In the meantime, there are numerous practical examples for the application of electrochemical noise measurements, e.g., in corrosion monitoring or in the quality control of incoming products.

Figure 3.113 shows a transient (localized dissolution of metal) as typically generated continuously during exposure of a stainless steel to a medium of approximately neutral pH which is free from chlorides. The amount of charge involved here corresponds to 1.5×10^{-12} C and is equivalent to approximately 5000 billion iron ions going into solution. This corresponds to a cubic defect with an edge length of ca. 40 nm for a face-centered cubic structure with a lattice constant of 0.364 nm. During the early stages, dissolution is crystallographically oriented and more hemispherical dissolution is only observed for much larger defects.

This tiny, active location very rapidly becomes covered again with a passive film, i. e., repassivation occurs. If one observes a freshly prepared surface on which the passive film is still being formed (Fig. 3.114a), it is possible initially to measure much larger transients. In the example shown, some 100 events greater than 25 pA were observed during the first 300 s. As time proceeds, the events become rarer and their amplitude decreases. The electrochemical noise behavior observed is significantly different, however, if the investigation is begun in a solution which already contains chloride ions (Fig. 3.114b). Firstly, many more events larger than 25 pA can be counted (ca. 300 here during the first 300 s). Secondly, the noise impulses decay only to a small extent, even after longer times. The effect of chlorides here can be understood not in terms of initiating local defects, but as delaying their repassivation.

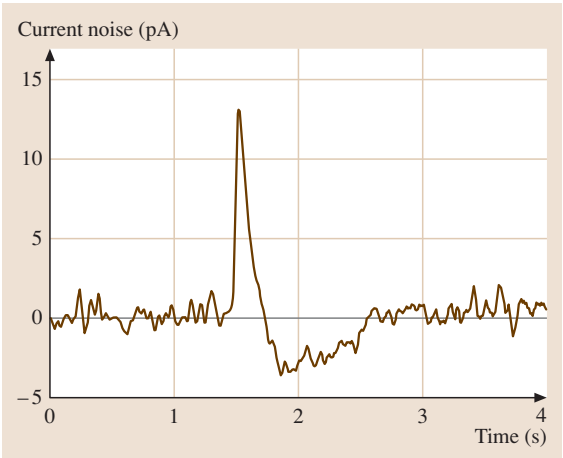


Fig. 3.113 Individual transient (maximum 13 pA) on a passive 18/10 Cr/Ni stainless steel in an oxygen-saturated aqueous solution free from chlorides (after [3.101])

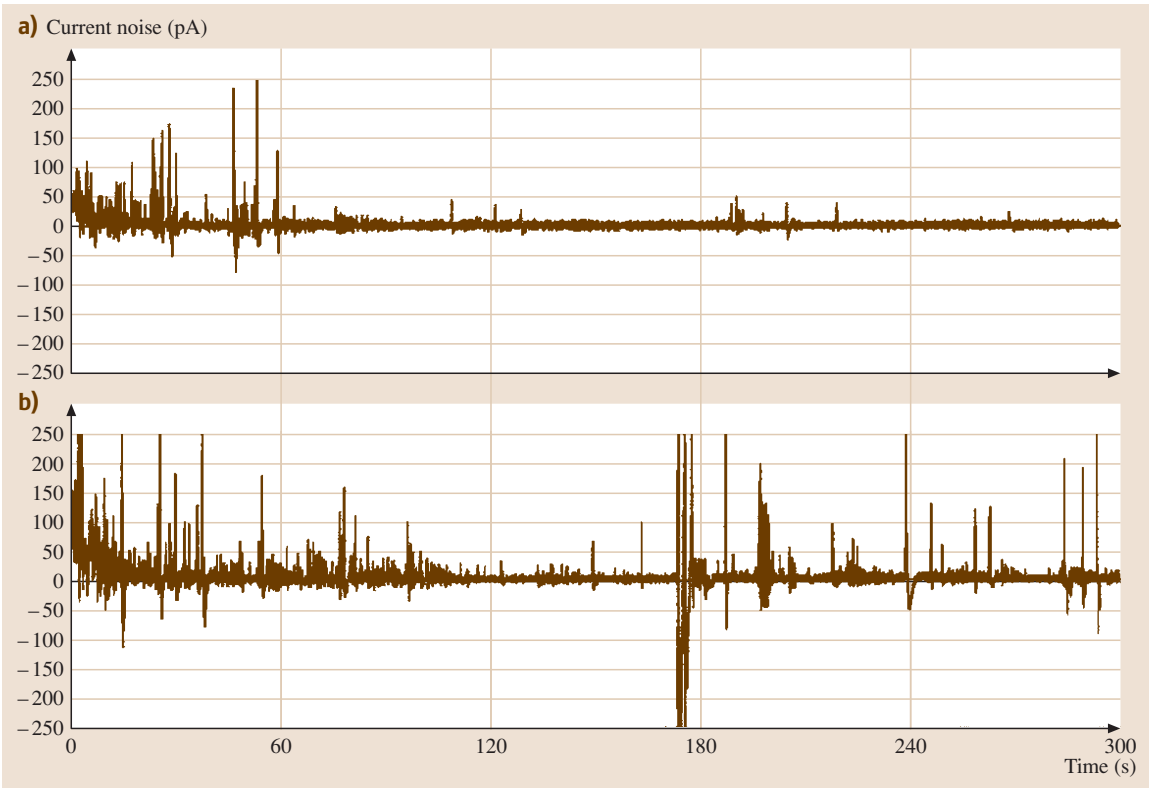


Fig. 3.114a,b Electrochemical current noise on a freshly prepared 18Cr/10Ni stainless steel surface in air-saturated aqueous solutions: **(a)** without Cl^-, **(b)**with 10^{-4} M NaCl (after [3.101])

It is thus incorrect to speak of chloride-induced corrosion, even though the behavior has this appearance. With larger amounts of chloride, and supported by impurities in the material, repassivation can be delayed to such an extent that the active locations become sufficiently large and stable to form the point of origin of corrosion pits (see later). In practice, this means that it is advisable to wait until the natural activity of passive film formation has virtually ceased before adding chloride ions when stainless-steel containers are filled. Depending upon the exact conditions, this may take hours or even days.

It is important for the fabricator or user of highly alloyed, stainless steels to be aware that he is not dealing with a robust material that is *inert* with regard to corrosion, but rather that the apparently *noble* properties of the steel arise from a very delicate film which needs to be treated with great respect. In essence, he needs to understand that something is involved which is invisible, but very effective, and that this can only function satisfactorily if the conditions remain conducive to film repair. If insufficient oxygen is present, as can occur, e.g., in crevices resulting from poor design, the rate at which the film reforms can be lower than the rate of dissolution, leading to rapid corrosion of the steel. Furthermore, the pervasive presence of chloride ions can delay local repassivation at spots where the film has *died*, or even prevent this completely. The latter case can lead to strongly localized dissolution of metal, i. e., pitting corrosion. Deposits of any kind that are present on the surface, whether visible or not, also make the formation of an intact passive film more difficult, or even impossible. For example, sweat from hand contact, dust, residues from tool abrasion, and fine rust particles can form the initiation points for corrosion which later becomes visible to the naked eye. Alterations to the metal itself, such as strong heat input, localized deformation, tensile stresses, etc., also influence the formation of the passive film and thus the corrosion behaviur.

In summary, even highly alloyed steels that are said to be resistant to rusting and to withstand exposure to acids are not in any sense *noble* and can corrode if they

are wrongly selected, stored, loaded, worked, welded, joined, combined, cleaned, pickled, ground, abraded, polished, or – more generally – insufficiently cared for. Furthermore, there is no point in trying to compensate for bad fabrication practices by choosing a more expensive steel, since such steels require special attention in the way they are treated and used in order for advantage to be taken of their inherently valuable properties. What has been stated here for highly alloyed, stainless steels is also valid for other metals and alloys which rely on passivation for their corrosion resistance. In every case, attention has to be paid (also during the design phase) to ensuring that repassivation can occur in an appropriate way.

Types of Corrosion

The different types of corrosion, and the associated damage they produce, are very varied. In addition to uniform surface attack, which is relatively straightforward, a number of nonuniform modes of attack often appear. These result from concentration elements (e.g., differential aeration cells), bimetallic couples, selective dissolution (e.g., intergranular corrosion), static and cyclic loads (stress corrosion cracking and corrosion fatigue), stray electrical currents, anions which hinder repassivation (pitting corrosion), rapid flow of the corrosion medium, etc.

Uniform Surface Attack. The progression of uniform surface corrosion can be predicted reasonably well. Thus, the loss of structural integrity as a result of corrosion reducing the wall thickness of a component can easily be compensated for by including extra margin during design. Note, however, that this only makes sense if the contamination of the medium from corrosion products can be tolerated. Even small amounts of heavy-metal ions can easily render a product unusable, or endanger the environment. Locations where dissolution occurs preferentially can result if protective layers are not formed uniformly, or if inhibitors are used at too low a concentration (dangerous attempts to protect against corrosion), and this leads to attack in the form of broad to sharp pitting.

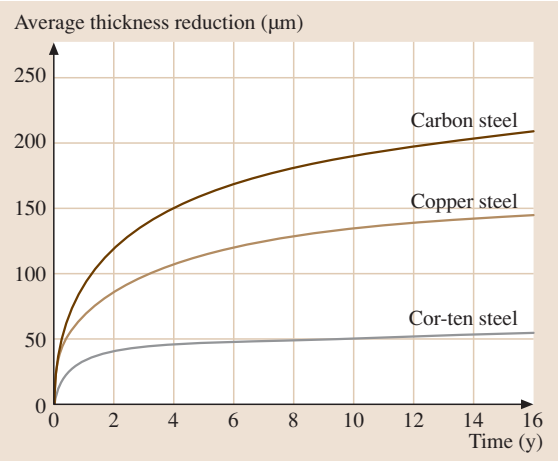


Fig. 3.115 Corrosion versus time curves for three steels in an industrial atmosphere (after [3.95])

The rate of corrosion is quoted in millimeters per year (mm/year) or in grams per square meter per year (g/m^2/year). If conditions for uniform surface dissolution occur, metals can be divided according to the velocity of lateral penetration (V_L) and the practical requirements into three main groups [3.103]:

1. $V_L \leq 0.15$ mm/year. The metals in this group have rather good corrosion resistance and are used for parts which would otherwise be especially endangered. These include, e.g., valve seatings, pump shafts and impellers, and springs.
2. $V_L = 0.15–1.5$ mm/year. The metals belonging to this group are adequate for service requirements where a higher rate of corrosion can be tolerated. Examples here include the fabrication of boilers, piping, valve bodies, bolt heads, etc.
3. $V_L > 1.5$ mm/year. Metals which corrode at such high rates are basically unsuitable for use in practice.

Steel corrodes in seawater, e.g., at a relatively uniform rate of around 2.5 g/m^2/year or 0.13 mm/year (not valid in tidal regions). Since the initial rate of corrosion is higher than the final value, such values are always averaged over a certain period of time. The measurement duration should always be quoted when making such statements.

Figure 3.115 shows the corrosion behavior versus time of three different steels exposed to an industrial atmosphere. The high initial rates are immediately apparent, as is the later improvement, particularly with Cor-ten steel, as a result of the formation of partially protective rust layers.

Flow-Induced Corrosion. All important technical metals are produced under the expenditure of a lot of energy and thus have a strong tendency to return to the lower-energy state by means of corrosion. They are prevented from doing so within reasonable limits by the existence of protective surface layers, which are themselves

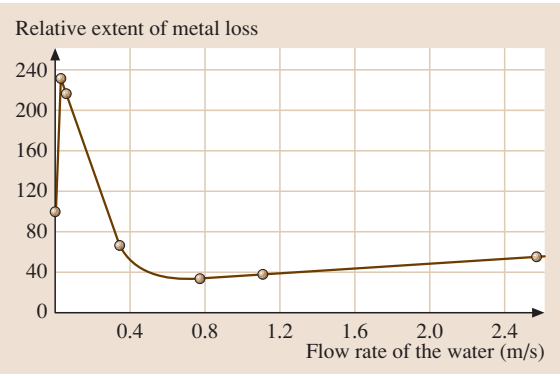


Fig. 3.116 Influence of the rate of flow of drinking water on the corrosion of an unalloyed steel (after [3.104])

produced via secondary reactions (usually involving oxidation). The flow rate of the medium is important in this context, as will be illustrated using the example of drinking water in a pipe made of unalloyed steel (Fig. 3.116). As the flow rate increases from zero, the corrosion rate initially rises, since more oxygen can be transported to the steel surface and thus more electrons can be consumed. At slightly higher flow rates, however, the corrosion products form a protective layer at the steel surface and the metal loss is significantly reduced. If the flow rate continues to increase, shear stresses eventually develop at the surface which lead to a reduction in thickness (and thus effectiveness) of the protective layer. The permissible flow rates are thus limited to a certain region. Table 3.10 gives guideline values for the water flow rate for various materials.

There is an additional consideration regarding the necessary oxygen concentration for formation of the protective layer and the velocity of water flow: at rates of around 0.5 m/s, 6 mg/l of dissolved oxygen is required, whereas only 2 mg/l are sufficient at rates above 1 m/s. The corrosion of highly alloyed, Cr/Ni stainless steels is not affected so much by flow rate. However, pitting corrosion occurs more readily in stagnant or weakly flowing water than when the medium is flowing rapidly. Each medium is associated with a particular type of protective layer. Thus if the medium changes, the layer also changes, or is dissolved and replaced by a different one. Frequent changes in the composition of the layer, however, may lead to complete loss of its protective nature, which can also happen as a result of temperature variations.

Table 3.10 Permissible velocities of water flow for various materials [3.104]

Material	v_A(m/s)[a]	v_{min}(m/s)	v_{max}(m/s)
Unalloyed steel	1.8	0.5	2.0
Galvanized steel	1.8	0.5	2.0
Steel with a duroplast coating	3.0	0.5	6.0
Cr/Ni stainless steel	4.8	0.5	5.0
Copper (99.7% pure)	1.0	0.7	1.2
Brass (CuZn 30)	1.8	1.0	2.0
Aluminum brass (CuZn20Al2)	2.3	1.0	2.5

[a] advisable velocity

Pitting Corrosion. Pitting corrosion is much feared with highly alloyed, Cr/Ni stainless steels, but also with titanium-, aluminum-, and nickel-base alloys. Under certain conditions, the repassivation rate (see *passivity*) is lowered to such an extent by factors such as high levels of halide ions, concentration of sulfides in the metal, lack of oxygen, etc. that the active locations become sufficiently large to be stabilized. They are then effectively decoupled from the bulk medium. Such pits often exhibit only a limited opening at the metal surface (thus reducing the diffusion of oxygen) and the internal electrolyte within the pit can become highly acidic (pH < 2) as the result of electrolytic processes. These conditions lead to extremely rapid local corrosion rates, since the fundamental reactivity of chromium is no longer restricted.

Pitting is affected not only by the nature of the metal surface, but also by the alloying elements. Apart from

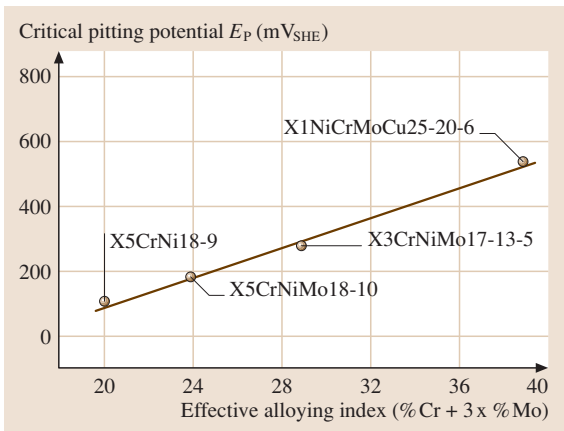


Fig. 3.117 Critical pitting potentials of various stainless steels as a function of their content of the alloying elements chromium and molybdenum (after [3.105])

chromium, molybdenum is especially important here. A so-called *effective index*, formed by multiplying the concentrations of individual alloying elements by appropriate factors, is used to describe this effect. This permits the creation of pitting resistance diagrams (plots of critical electrochemical pitting potential versus effective alloying index), as is shown in Fig. 3.117 for various common steels.

The critical pitting temperature (CPT) is also often used to describe pitting behavior, since temperature is a key factor affecting this type of corrosion. The advantage here is that the determination of this value can be made without the need for external instrumentation (such as an electrochemical potentiostat). The temperature of a 10% solution of $FeCl_3$ is raised every 24 h by 2.5 °C until pitting corrosion finally becomes visible. Such a CPT curve also goes up as the effective alloying index is increased. It is possible to determine critical pitting temperatures more quickly (in about 30 min) by the measurement of electrochemical noise.

Since the passive film can always be damaged in practice, it would really be more important to know whether such defects can be successfully repaired, rather than if pits are formed. Unfortunately, however, the determination of so-called repassivation potentials (or repassivation temperatures) is presently too inaccurate to permit their use in routine testing. Nevertheless, efforts continue to determine such values more exactly, because of their fundamental importance.

Intergranular Corrosion. Precipitation of a new phase occurs in a mixed crystal lattice if the solubility of one of the components is exceeded. This occurs preferentially at the grain boundaries, for thermodynamic reasons (lower energy). If the phase concerned is less corrosion resistant, and if it forms a continuous network, dissolution takes place particularly at the grain boundaries. In some cases, this can lead to the total decomposition of the material into individual crystals (grains), as can be the case, e.g., with the Al_3Mg_2 phase in aluminum/magnesium alloys, or the less noble β-phase in brasses.

When highly alloyed, Cr and Cr/Ni stainless steels are reheated (e.g., during welding), precipitation of chromium carbides ($Cr_{23}C_6$) can occur. The formation of this phase rich in chromium (up to 85%) leads to chromium depletion in the immediate vicinity (Fig. 3.118). In acid media, no stable passive film is formed if the chromium level sinks below a critical value of around 12%, which then results in extremely high rates of dissolution (up to 1 million times higher

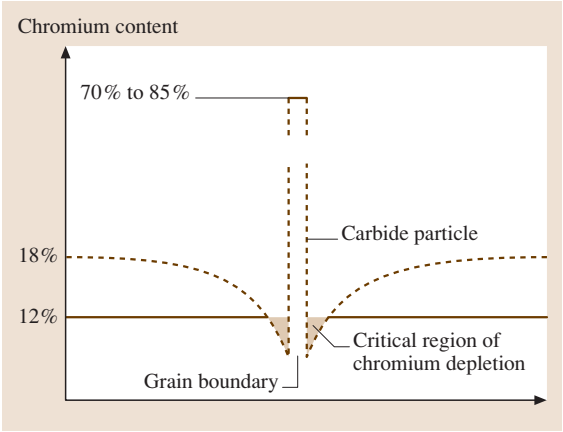


Fig. 3.118 Schematic representation of the distribution of chromium at a grain boundary in a sensitized stainless steel containing 18% chromium (after [3.106])

at the grain boundary than in the interior of the grains).

Figure 3.119 shows a photograph taken with an atomic force microscope (AFM). Strong dissolution can be seen to have taken place at the grain boundaries. The light-colored, protruding particles are the carbides which have been left behind.

The formation of a passive film is very dependent upon the electrochemical corrosion potential, which, in turn, is strongly influenced by the pH value of the solution. This results in an apparent paradox with passive steels, as a sensitized material can undergo severe intergranular corrosion in weakly acid media (beer, wine, hair shampoo) but exhibit little or no corrosion in a much more acid environment, where the potential is displaced to more positive values. In strongly oxidizing acids, however, where the potential is even higher, dis-

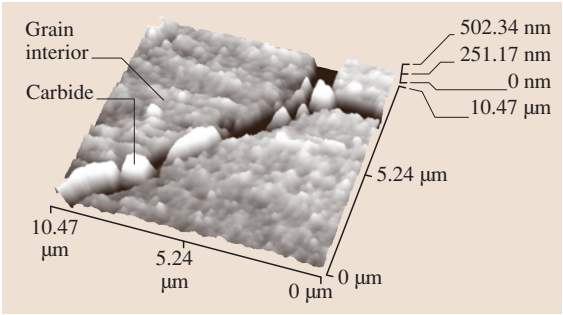


Fig. 3.119 AFM photograph of a highly alloyed stainless steel undergoing intergranular corrosion, leaving behind chromium-rich carbides (after [3.107])

solution of the chromium carbides themselves can take place, again leading to intergranular corrosion.

Stress–Corrosion Cracking. Stress corrosion cracking is a very dangerous form of attack, since it often leads to component failure without visible signs of damage at the metal surface. The preconditions for it to occur include a susceptible metal, a specific aggressive medium, and a critical level of stress. Internal stresses, such as that caused through prior cold working or heat treatment, can contribute to the required loading. As with pitting corrosion, stress corrosion cracking is observed in materials which are otherwise protected by surface layers. Those affected include unalloyed, low-alloy, and highly alloyed steels, as well as nickel-based alloys, aluminum, and brasses. A distinction is made between anodic and cathodic stress corrosion cracking, although mixed types are also often observed.

With anodic stress corrosion cracking, crack formation arises from a dissolution process, whereby localized defects, slip bands, phases susceptible to corrosion, and grain boundaries can all play an important role. Often, the crack forms at an incipient corrosion pit and the progress of corrosion involves alternating phases of crack propagation and broadening of the crack through dissolution. Localized embrittlement of material can occur ahead of the crack tip. Figure 3.120 shows schematically one example of the many possibilities leading to cracking.

The formation of atomic hydrogen plays a decisive role in cathodic stress corrosion cracking. This originates from the cathodic reaction according to (3.4), whereby compounds of sulfur, in particular, prevent the hydrogen atoms from recombining to molecular hydrogen. These hydrogen atoms are easily able to enter the metal and then form hydrogen gas, which cannot diffuse easily, or metal hydrides. In this way, dislocation movement is blocked and the metal becomes locally brittle (hydrogen embrittlement). Cracking of the material can then take place above a critical stress level.

The crack path in both types of stress corrosion cracking can follow the grain boundaries within the microstructure (intergranular), or propagate through the interior of the grains (transgranular), independent of the specific causes. Cathodic stress corrosion with transgranular cracking plays a special role with steels of higher strength. Without hydrogen, these would withstand much higher levels of stress, so that this particular type of corrosion often ultimately limits their use.

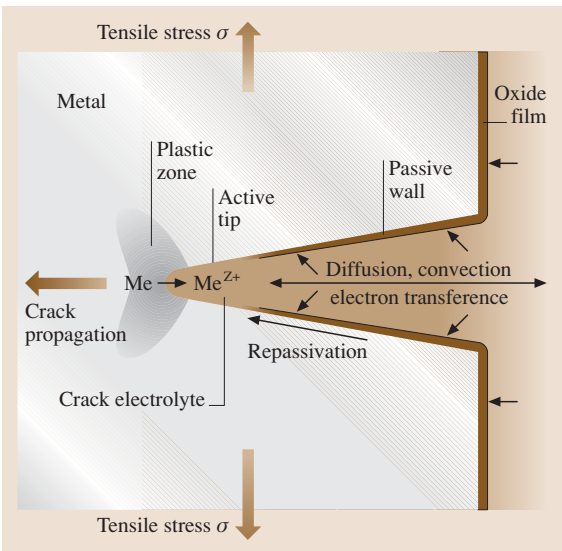


Fig. 3.120 Modeling of stress corrosion crack propagation in a passive metal by quasicontinuous, intermittent crack-tip-slip activation followed by deformation-enhanced dissolution of metal at the crack tip. Note the role of a plastic deformation zone at the crack tip and protection of the crack walls by repassivation, after crack growth has occurred (after [3.108])

Corrosion Fatigue. Corrosion fatigue can occur if a corrosion process occurs at the same time as cyclic mechanical loading. The combined effect is easily assessed in terms of the relevant *S*–*N* curve (Fig. 3.121).

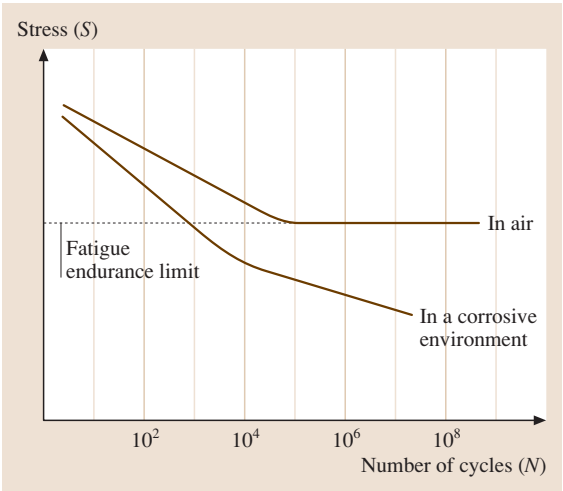


Fig. 3.121 Alteration of the cyclic *S*–*N* curve as a result of corrosion fatigue (after [3.106])

A corrosive environment leads to the absence of a true fatigue endurance limit. Instead, the fatigue strength can only be stated as a function of time (and accumulated loading cycles). The initial process of crack formation is comparable to that occurring in a non-corrosive environment: elements of the lattice structure become separated from the surface at slip bands as a result of localized plastic deformation. This results in the formation of microscopic notches, leading to stress concentrations, and later to cracks. In a corrosive environment, however, the cracks propagate more quickly. As a rule, they are transgranular in nature. All materials are basically affected and no specific corrosive medium is required. The damage results from the slip processes that are initiated by cyclic loading.

Erosion Corrosion. If the corrosion of metallic materials is stimulated by erosion processes at the metal surface, the damage mechanism is referred to as erosion corrosion or cavitation. Erosion corrosion can be observed in equipment containing flowing water, or steam, as a result of high flow rates and the presence of solid particles in the medium. The latter damage the microstructure by impacting the metal surface and thus input mechanical energy, which favors corrosion.

Cavitation corrosion refers to the situation when gas contained in water is abruptly released, or transformed into steam. The collapse of the resulting bubbles damages the metal surface by releasing soft or brittle components from the microstructure, thus stimulating the corrosion process. Cavitation corrosion is observed, particularly in steam boilers, degassing equipment, pumps, turbines, and valves.

Galvanic Corrosion. In practice, an attempt is often made to explain all corrosion phenomena by reference to the list of standard electrode potentials. However, the theory of galvanic corrosion elements derived from this has been unacceptable scientifically since the investigations of *Wagner* and *Traut* in 1938 [3.109]. It should be regarded only as a special case of the more universal theory of mixed potentials.

So-called galvanic corrosion occurs, in addition to normal corrosion, if two metals with different electrochemical potentials are connected together electrically. In this case, metal dissolution is accelerated at the less noble material (anode) and the consumption of electrons is favored at the more noble material (cathode). It is impossible to say what will be more or less noble just from the list of standard electrode potentials, since the addition of alloying elements and the formation of protective surface layers result in an entirely different order.

Table 3.11 Influence of area ratio on the corrosion rate of shiny nickel in contact with chromium in simulated rainwater of pH 2.5 (the less noble chromium, according to the list of standard potentials, forms the cathode here and is nobler than nickel as a result of passive film formation) [3.96]

Area ratio Cr/Ni for constant chromium area of 6.3 cm^2	Anodic current density of nickel dissolution (mA/cm^2)	Rate of nickel metal loss (mm/year)
1 : 1	0.0015	0.016
1 : 0.1	0.015	0.16
1 : 0.01	0.15	1.6
1 : 0.001	1.3	13.9
1 : 0.0001	6.8	72.8
1 : 0.00005	17	182

In practice, the contact resistances and the conductivity of the electrolyte are often more important than the potential difference. The area ratio of anode to cathode is also of great importance. Table 3.11 shows the effect of area on the current density using, as an example, passive chromium as the cathode and active nickel as the anode. From this it can be seen that the anode should be as large as possible and the cathode as small as possible. In practice, aluminum sheets (large anode) can be joined together with Monel rivets (70% Ni, 30% Cu) without leading to problems of galvanic corrosion. If one were to join copper sheets with aluminum rivets (small anode), however, the results would be catastrophic.

Microbiologically Influenced Corrosion. Corrosion caused by bacteria has increased in importance over recent years. Thus, damage to materials in the Earth (e.g., pipes and cables) has occurred as a result of the effects of micro-organisms (microbiologically influenced corrosion, MIC). One such example involves corrosion processes as a result of sulfate-reducing bacteria: in the presence of water, these can reduce sulfates and simultaneously lower the pH value with the formation of sulfuric acid. Traces of water are contained even in fuels such as oil and p.t.o., so that microbes can develop and disturb the electrochemical equilibrium. The resulting electrochemical reaction releases oxygen and thus permits electron consumption, leading to notch-like defects at the surface of the material. Although the suspicion is often raised that the bacteria themselves are directly active (*iron eaters*), this is not true. Instead, the attack is related to digestive products (e.g., acids), as well as to hindered access of the oxygen necessary for repassivation resulting from the formation of microbe colonies

Table 3.12 Influence of prior surface preparation on the lifetime of an alkaline-epoxy-based coating (consisting of one primer, two intermediate, and one final layers) exposed outdoors [3.96]

Prior surface and manner of surface preparation	Average lifetime of the coating system
Rust	1–2 years
Converted or stabilized rust	1–3 years
Scale (firmly adherent)	3 years
Manual derusting	4 years
Prepared with mechanical tools	5 years
Flame descaled	5 years
Pickled	8–10 years
Blasted	9–12 years

at the surface of the material. Clarification of the exact corrosion mechanism in an individual case can be complicated, since one is dealing with a living system and the local conditions can vary considerably (aerobic or anaerobic bacteria).

Corrosion under Coatings. The corrosion mechanism under coatings is still somewhat unclear and research is still needed into the effects of a series of influencing factors. As a rule, coatings are hydrophobic, i. e., water droplets do not wet the surface. This is only valid, however, for liquid water, where thousands of molecules band together to form small clusters. Although invisible, water vapor (not to be confused with steam, which also contains clusters) consists of separate molecules and determines the relative air humidity. Such water molecules can diffuse relatively easily through a coating, as can oxygen, carbon dioxide, and sulfur dioxide. If the coating adhesion is poor, cavities (or even rust particles) can exist between the metal surface and the coating and these permit local condensation of water and concentration of metal ions. Together with the water, the oxygen which diffuses into such cavities initiates the electron-consuming process with the formation of OH^- ions. These combine with the iron ions which have gone into solution to form rust. Since porous rust has a volume which is six to eight times greater than that of the corroded amount of metal, the coating is pushed away from the surface (formation of blisters). Larger amounts of water then collect in the resulting cavities and accelerate the processes already described. Table 3.12 illustrates the life expectancy as a function of the preparation of the surface prior to coating.

This makes it clear that the lifetime can be very different, even for the same coating system. With a firmly adhering coating, the water molecules still obtain access to the surface very quickly, but the locations at which they can condense remain so small that changes only become visible to the naked eye much later.

3.6.3 Corrosion (Chemical)

Basic Principles

With chemical corrosion, the material and the medium react directly with one another as a result of an overlap being formed between the electron paths of each of the partners. No increase in free electrons occurs in the metal. The products formed determine the continued evolution of the corrosion. The formation of protective layers is also desirable here, since these layers act as effective barriers to diffusion processes and, thus, hinder further reactions. The extent of corrosion can be determined either gravimetrically (weight change) or metallographically.

In contrast to the above, electrochemical corrosion leads to processes which take place in parallel at separate locations. Corrosion products (rust) are formed via secondary reactions, i. e., after the actual corrosion has occurred. The free electrons which are generated offer the possibility of direct measurement of the corrosion processes involved.

High-Temperature Corrosion

At high temperatures, the corrosion resistance of metallic materials decreases as a result of reactions with gases. The reaction product here is referred to as scale. It is a solid corrosion product which grows at the metal surface and forms a barrier to the reaction partners metal and gas. In order for this layer to grow, at least one of the partners must be mobile within the layer. Many oxides and sulfides contain cavities and vacancies within their microstructure and these locations permit metal cations to be transported towards the outside.

Scale formation is particularly important in practice with steels which are exposed to oxygen from the air, or to mixtures of common technical gases with steam or carbon dioxide. At low temperatures (200–400 °C), the initially high rate of reaction rapidly falls to very low values and growth of the protective layer versus time can be described by a logarithmic equation. In general, the resulting thin films ($< 0.1\,\mu m$), which are often described as tarnish layers, do not represent any significant damage to the material. They can, however, be detrimental upon subsequent exposure to water, i. e., in connection with electrochemical corrosion. At higher temperatures, the initial chemical reaction involves the

formation of thicker layers, free from pores, which grow with time according to a parabolic law. In this case, the diffusion speed of the ions and electrons is rate limiting. If the coverage is incomplete, however, as a result of the formation of pores and cracks, then either the reaction at the phase boundary metal/gas, or the supply of oxygen, become rate limiting. In these cases, layer thickening occurs in a linear manner with time, i. e., the metal is progressively destroyed (catastrophic corrosion). In oxygen or air at temperatures above 570 °C, iron forms a complex scale involving the following layers $Fe-FeO-Fe_3O_4-Fe_2O_3-O_2$. The proportion of wüstite (FeO) amounts to almost 90%, whereas magnetite (Fe_3O_4) represents 7–10% and the haematite layer (Fe_2O_3) only 1–3%. Figure 3.122 shows the relevant phase-stability diagram for iron and oxygen.

If slow cooling occurs below 570 °C, wüstite decomposes into iron and magnetite. The resulting layers are brittle and full of microcracks, both because of the differing density and the relative lack of *ductility* of Fe_3O_4 in comparison with FeO. Rapid cooling, as occurs, e.g., during hot-forming of steel plates, prevents this transformation and the resulting scale remains adherent. The oxidation rate of steels can be decreased by alloying with chromium, aluminum, and silicon.

Another set of damage mechanisms exists and leads, e.g., to carburization or decarburization of steels. Thus exposure of Cr/Ni stainless steels to gas atmospheres which can supply C results in carbide formation, which

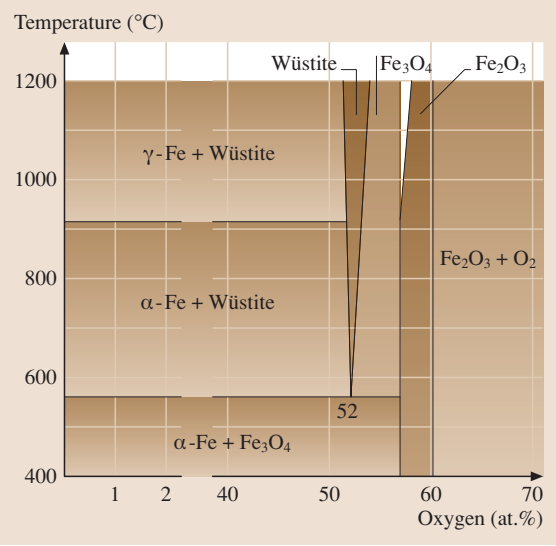


Fig. 3.122 Phase stability diagram for iron and oxygen (after [3.110])

continues to form within the body of the steel. Internal carburization is particularly detrimental for toughness at low temperatures. Particularly catastrophic carburization is possible for steels within an intermediate temperature range (400–600 °C) and is described as *metal dusting*. The material is transformed into a fine powder consisting of metal and carbon. After rapid oversaturation of the material with carbon, the process starts with the formation of an unstable carbide Me_3C (Me=Fe, Ni) at the surface and at grain boundaries. This is followed by decomposition of the carbide according to

$$Me_3C \rightarrow 3Me + C\,. \qquad (3.91)$$

The resulting, fine particles of metal act in a catalytic manner to accelerate the further uptake of carbon, so that voluminous carbon deposits grow on the metal surface. These loose deposits can be removed by the gas stream, leaving behind indentations resembling pitting corrosion.

Damage as a result of decarburization can occur in plants using pressurized hydrogen for the purposes of synthesis. Atomic hydrogen, formed as a result of thermal dissociation above 200 °C, becomes dissolved in the steel and reacts with iron carbides, producing methane

$$Fe_3C + 4H_{(Fe)} \rightarrow CH_4\,. \qquad (3.92)$$

This gas cannot escape, because of its molecular size, and leads to the build-up of high internal pressures in the metal. The mechanical properties of the steel are negatively affected by decarburization and embrittlement, with the result that inclusions, grain boundaries, and similar material separations can form the initiating points for brittle fracture.

In copper which contains oxygen, hydrogen reacts with copper oxide to form steam

$$Cu_2O + H_2 \rightarrow 2Cu + H_2O\,. \qquad (3.93)$$

This results in pores, which can become joined together to form networks of cracks. Such damage is known as *hydrogen sickness*.

More information regarding high-temperature oxidation can be found, e.g., in [3.111].

Corrosion of Glasses

Nonmetallic, inorganic materials are relatively resistant to attack at room temperature in most organic and inorganic solutions, in water, and in acids or weak bases. The extent of their resistance is dependent upon their chemical composition, the material microstructure, and

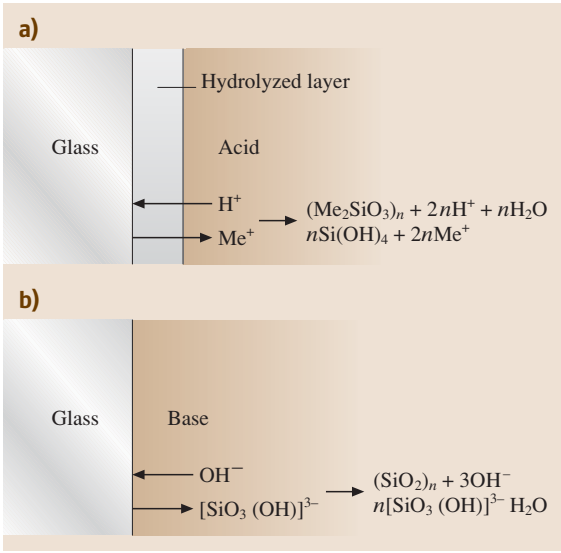


Fig. 3.123a,b Corrosion of glass by (a) acids, (b) bases (after [3.112])

the environmental conditions. Glass has no uniform composition, although it may appear relatively homogeneous from the outside. It contains various microscopic phases of different composition and is crystalline to different extents. Desired properties can be achieved by influencing the form and condition of these phases in a specific way. Thus normal glass, containing as its main component SiO_2, can be significantly attacked by hydrofluoric acid. However, special types of glass on the basis of P_2O_5 and Al_2O_3 exhibit good properties even in this medium.

The resistance of normal types of glass and enamel in acid media (apart from hydrofluoric acid) arises from the fact that the hydrogen ions in the acid are exchanged with migrant cations (Na^+, K^+, Li^+, Ca^{2+}). As this exchange proceeds, hydrolysis leads to the formation of a gelatinous layer rich in silica which limits the diffusion of ions, thus hindering corrosion and making the glass increasingly resistant to attack. This gelatinous layer only becomes damaged and allows further attack on the glass under very unfavorable conditions, e.g., during exposure to superheated steam or in heat exchangers, where additional chemical reactions involving carbon dioxide from the air can occur. Figure 3.123 shows schematically the mechanism of glass corrosion.

The resistance of the glass is very much lower in alkaline solutions (bases), since the OH^- ions can destroy the Si–O–Si links through chemical reactions producing low-molecular silicates, which dissolve in the medium. The loss of material follows a linear progression with time and can be appreciable, particularly in strong bases, at temperatures above 30 °C.

Even rainwater can lead to detectable corrosion of glass after long periods of time. Figure 3.124 shows the surface of an approximately 50-year-old pane of glass at high magnification. No amount of intensive cleaning can return the shine to this window.

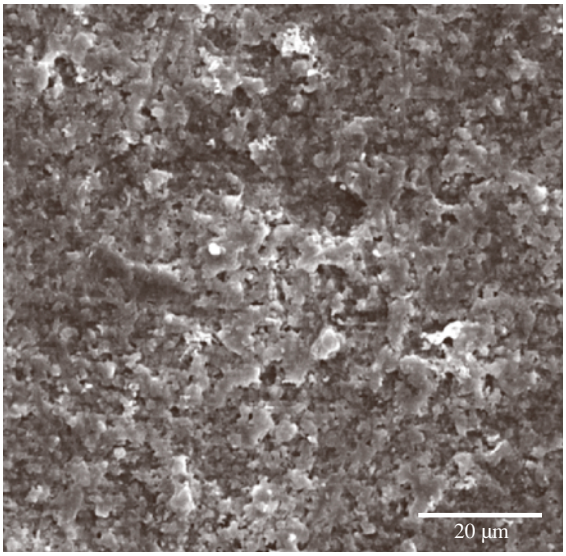


Fig. 3.124 Surface of a pane of glass as seen in the electron microscope after approximately 50 years of exposure outdoors

Corrosion of Polymers

Polymers are essentially resistant to attack in media such as the atmosphere, aqueous solutions, acids, and bases in which most metallic materials corrode. However, this does not mean that polymers show no corrosion in general. They lack resistance to attack in various organic solvents. Furthermore, they can exhibit damage due to corrosion in other media, the extent of which depends both on the chemical composition and structure of the polymer, as well as on the concentration of reactant in the medium concerned, the temperature, and the exposure time. In contrast to metals, the corrosion of plastic almost always begins with the entry of foreign molecules, i. e., with a physical process occurring in three steps: adsorption (wetting of the surface by the corrosive medium), diffusion (entry of the medium into the material), and absorption/swelling (uptake of the medium with uniform and complete penetration of

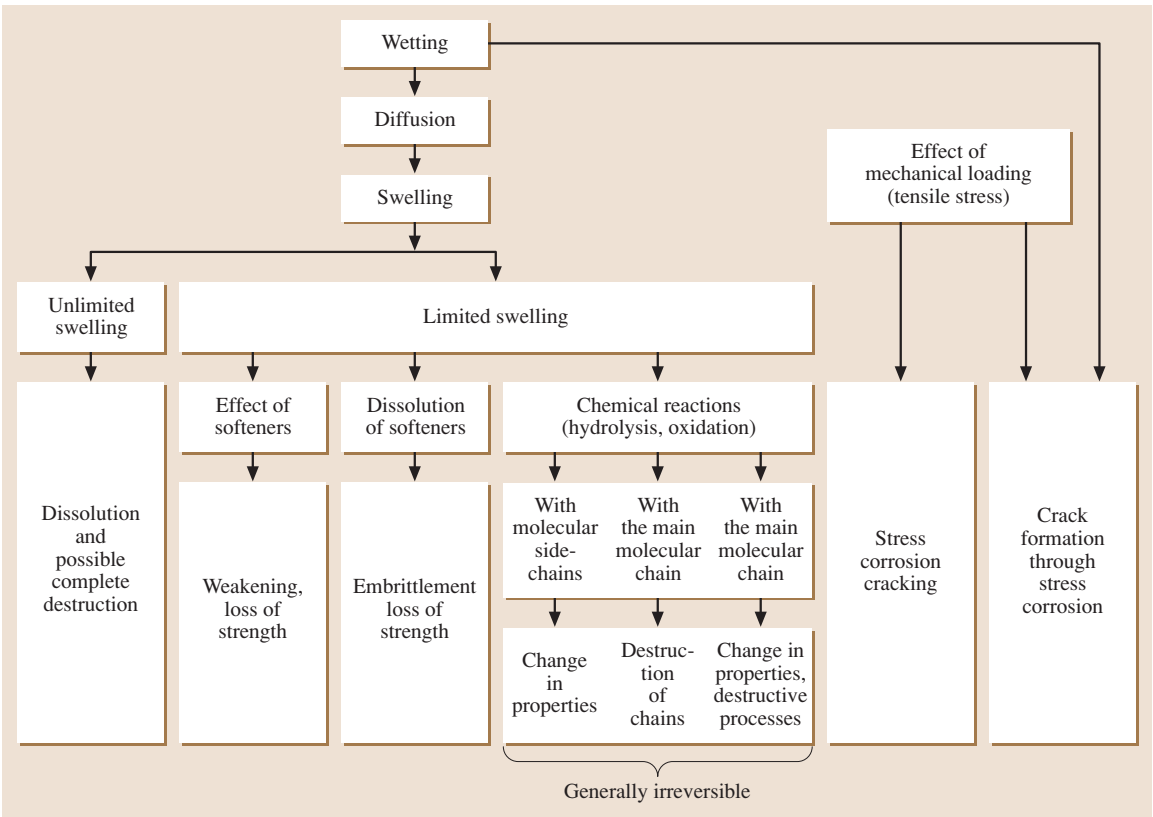


Fig. 3.125 Damage caused to plastics by liquid media

the material). Figure 3.125 shows the individual steps in the corrosion of polymers.

After the initial steps, chemical processes (chemisorption, oxidizing or reducing attack, hydrolysis, etc.) can occur and lead to considerable deterioration of the properties of the material. Attack on plastics usually occurs in a complex manner and is accompanied by further damage, such as the attainment of thermodynamic equilibria (late crystallization, recrystallization, relaxation of stress, and deformation). This takes place under the influence of thermal and/or radiation energy, as well as through biological effects. The irreversible change in properties which then occurs progressively with time is usually referred to as ageing. As with metals, stress corrosion cracking is also possible with plastics. Its occurrence again requires the presence of a specific, aggressive medium, as well as internal and/or external tensile stresses.

3.7 Materials in Mechanical Engineering

Engineering materials, in principle, may be divided into four main classes:

1. Metals
2. Ceramics and glasses
3. Polymers and elastomers
4. Composites

Materials belonging to one of these classes exhibit comparable properties, processing routes, and most often applications as well. The criteria for the material selection are rather complex and depend on the intended application purpose. To the main design criteria belong strength, stiffness, fracture toughness, formability, joinability, corrosion resistance, coefficient of thermal

Table 3.13 Properties of some widely used metallic materials, carbon fiber, and high-density polyethylene (HDPE). Note that some of the values given in the table are prone to variation (data compiled from different sources [3.115–117])

Metal	Melting point (°C) base metal	Density (g/cm^3)	Yield strength (MPa)	Specific yield strength ($MPa\,cm^3/g$)	Young's modulus (GPa)	Cost (US$/t)
High-carbon steels	1536	7.8	350–1600	45–205	210	200
Stainless steels	1536	7.8	150–500	19–64	193	2700
Cast irons	1147 (eutectic)	7.4	50–400	7–54	150	160
Aluminum 2000 series	660	2.8	200–500	71–179	70	1430
Titanium alloys	1668	4.5	400–1100	89–244	100	6020
Copper alloys	1083	8.9	75–520	8–58	135	1330
Superalloys	1453	7.9	800	101	180	6500
Magnesium alloys	650	1.75	300	171	45	2800
Carbon fiber	3650	1.75	3500–5500	2000–3140	230–400	30 000
High-density polyethylene (HDPE)	~ 250	0.95	26–33	27–35	0.7	1000

expansion, cost, and last but not least recyclability. For structural applications in mechanical engineering metallic materials [3.113, 114] are still the most widely used group of materials; their order of importance is Fe, Al, Cu, Ni, and Ti. While the physical properties of materials belonging to different classes are given in Sect. 3.3, in Table 3.13 a comparison of the mechanical properties of some important metals and alloys, carbon fiber, and a polymer is shown.

3.7.1 Iron-Based Materials

Iron-based materials are the most widely used metallic materials, mainly because of their relatively inexpensive manufacturing and their enormous flexibility. Accordingly, the properties of Fe-based materials can be varied to a great extent, allowing precise adaptation to specific application requirements ranging from high-strength, high-temperature, and wear-resistant alloys for tools to soft or hard ferromagnetic alloys for applications in the electrical industries.

Pure iron, however, is only of minor importance in structural applications since its mechanical properties are simply inadequate. Alloying with carbon leads to the most important groups of constructional alloys, namely:

1. Steels with a carbon content of up to about 2.06% carbon (if not stated otherwise all compositions are giving in wt. %)
2. Cast iron, which practically contains 2.5–5% carbon

These Fe−C alloys exhibit outstanding properties, including widely variable mechanical properties: yield strengths ranging from 200 MPa to values exceeding 2000 MPa, hot and cold rolling ability, weldability, chip-removing workability, high toughness, high wear resistance, high corrosion resistance, heat resistance, high-temperature resistance, high Young's modulus, nearly 100% recyclability, and many more.

In the following sections the characteristic phases, microstructures, compositions, and applications of iron–carbon alloys are treated with emphasis on the fundamental background. For further reading, references such as [3.1, 118–122] and the online database [3.123] are recommended.

The Iron–Carbon Phase Diagram and Relevant Microstructures

Fe−C-based materials, in general, can be classified into two main categories:

1. Steels or steel castings, which are forgeable iron–carbon alloys with up to about 2.06% C
2. Gray iron or pig iron with more then 2.06% C (in practice 2.5–5%), which cannot be forged and are brought into final form only by casting

These two groups of Fe–C alloys divide the iron–carbon diagram (Fig. 3.126) into two parts, namely an eutectoid (steel) part and an eutectic (cast iron) part. In the thermally stable condition carbon prevails in the

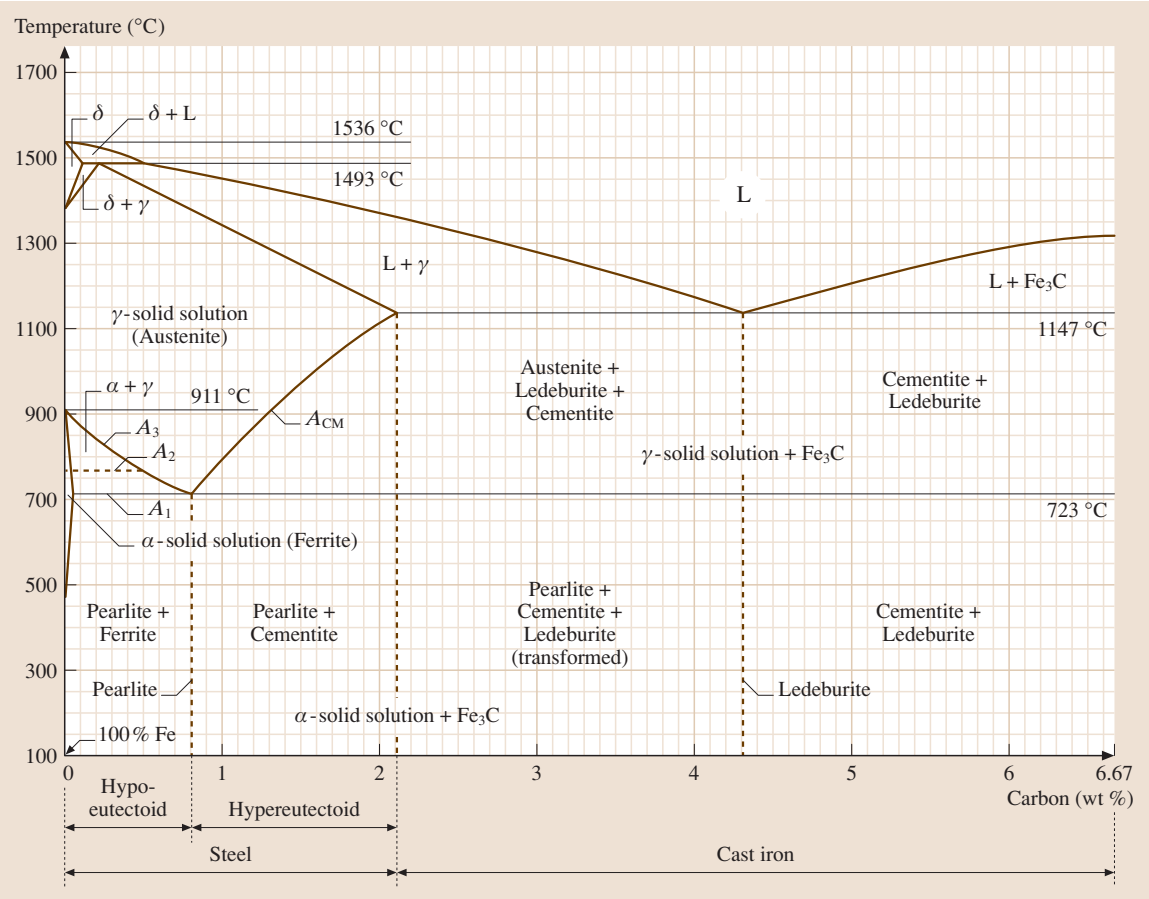


Fig. 3.126 The metastable Fe−Fe_3C (6.67%C) diagram

form of graphite. Although graphite or more precisely its shape and proportion plays a major role in adjusting the properties of cast irons, this equilibrium phase is usually not obtained in common steels. Instead, carbon in steels emerges in the form of metastable iron carbide (Fe_3C-cementite). Therefore, the metastable equilibrium (Fig. 3.126) between iron and iron carbide is relevant to the behavior of most steels in practice. A closer look at the Fe–Fe_3C phase diagram reveals the three fundamental ideal diagrams introduced in Sect. 3.1.2, namely a peritectic and a eutectoid system in the steel part and a eutectic system in the cast-iron part of the diagram.

Pure iron appears in three different allotropic forms for which the following notations are used:

- α-Fe with the bcc structure, which is stable at temperatures below 911 °C – note that from 769 °C (A_2 line) to lower temperatures α-Fe is ferromagnetic without a lattice transformation.
- γ-Fe with the fcc structure, which is stable between 911 °C and 1392 °C.
- δ-Fe with the bcc structure, which exists from 1392 °C to the melting point at 1536 °C.

With its significantly smaller atomic radius carbon occupies the interstitial lattice sites (compare Sect. 3.1.2) of the iron phases. The solubility, however, depends on the size of the lattice gap and therefore on the lattice type of the specific Fe phase (compare Fig. 3.126). These differences in the maximum solubility of carbon are the basis for the enormous variability of the mechanical properties of steels. In Table 3.14 phases and phase mixtures of the Fe–Fe_3C system, their corresponding (maximum) carbon content at different temperatures,

Table 3.14 Phases and phase mixtures of the Fe–Fe_3C diagram (SS: solid solution)

Phase and phase mixture	Maximum carbon content (in single-phase region) or percentage of particular phase in phase mixtures	Designation
α-Fe	0.02% (at 723 °C)	Ferrite
γ-Fe	2.06% (at 1147 °C) 0.8% (at 723 °C)	Austenite
δ-Fe	0.1% (at 1493 °C)	δ-Ferrite
Fe_3C	6.67%	Cementite
(α-Fe + Fe_3C)	88% α-SS + 12% Fe_3C	Pearlite (eutectoid)
(γ-Fe + Fe_3C)	51.4% γ-SS + 48.6% Fe_3C	Ledeburite I (eutectic)
(α-Fe + Fe_3C)	35.5% α-SS + 64.5% Fe_3C	Ledeburite II

and their microstructural nomenclature are summarized.

The intermetallic compound Fe_3C (*cementite*) or more accurately their microstructural appearance plays a crucial role for the adjustment of the mechanical properties of steels. Cementite with 6.69% carbon is based on an orthogonal lattice where dislocation glide at low temperatures is nearly impossible. It exhibits therefore a very high hardness (1400 HV) and brittleness. However, in the form of finely distributed particles or lamellas in the grain interiors it can hinder dislocations from glide very effectively. At a carbon content of about 0.8% at which the eutectoid reaction occurs a phase mixture of 88% α-Fe and 12% Fe_3C (eutectoid) is formed from γ-Fe(C) solid solution (compare Table 3.14). The typical arrangement of the eutectoid in contiguous lamellae (Fig. 3.127) is the result of fast decomposition and the designation *pearlite* is used for this microstructure.

Likewise the eutectic microstructure at about 4.3 wt. % C, i. e., a phase mixture of 51.4% γ-Fe and 48.6% Fe_3C, is called *ledeburite*. At carbon concentrations which vary from the exact eutectoid or eutectic composition, the microstructure contains more than one component. This is shown in Fig. 3.128 for *hypo-* and *hyper*-eutectoid steels.

On slow cooling of hypo-eutectoid compositions, i. e., of alloys containing less than 0.8% C, the austenite partly transforms to ferrite in the temperature range of 911–723 °C. Since the solubility of carbon in α-Fe is significantly lower than in γ-Fe the residual austenite simultaneously enriches in carbon along the A_3-line, until at 723 °C the remaining austenite, now exactly at the eutectoid composition, transforms to pearlite as a second microstructural component (Fig. 3.128). Hypereutectoid alloys with 0.80–2.06% carbon first form cementite at the γ-Fe grain boundaries on cooling in the temperature interval 1147–723 °C, while the austenite depletes in carbon. The carbon concentration of the austenite reaches in turn 0.8% at 723 °C and it transforms to pearlite on further cooling. The microstructural composition of Fe–C alloys in the thermodynamic

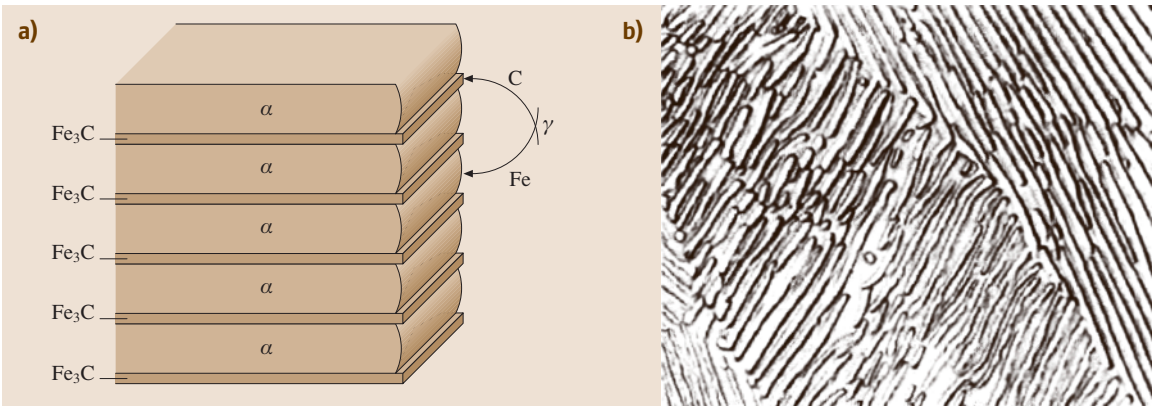


Fig. 3.127a,b Transformation of austenite to pearlite below 723 °C. **(a)** Decomposition of γ-Fe into lamellas of two different phases (α-Fe and Fe_3C). **(b)** Microstructure of pearlite lamellas (after [3.54])

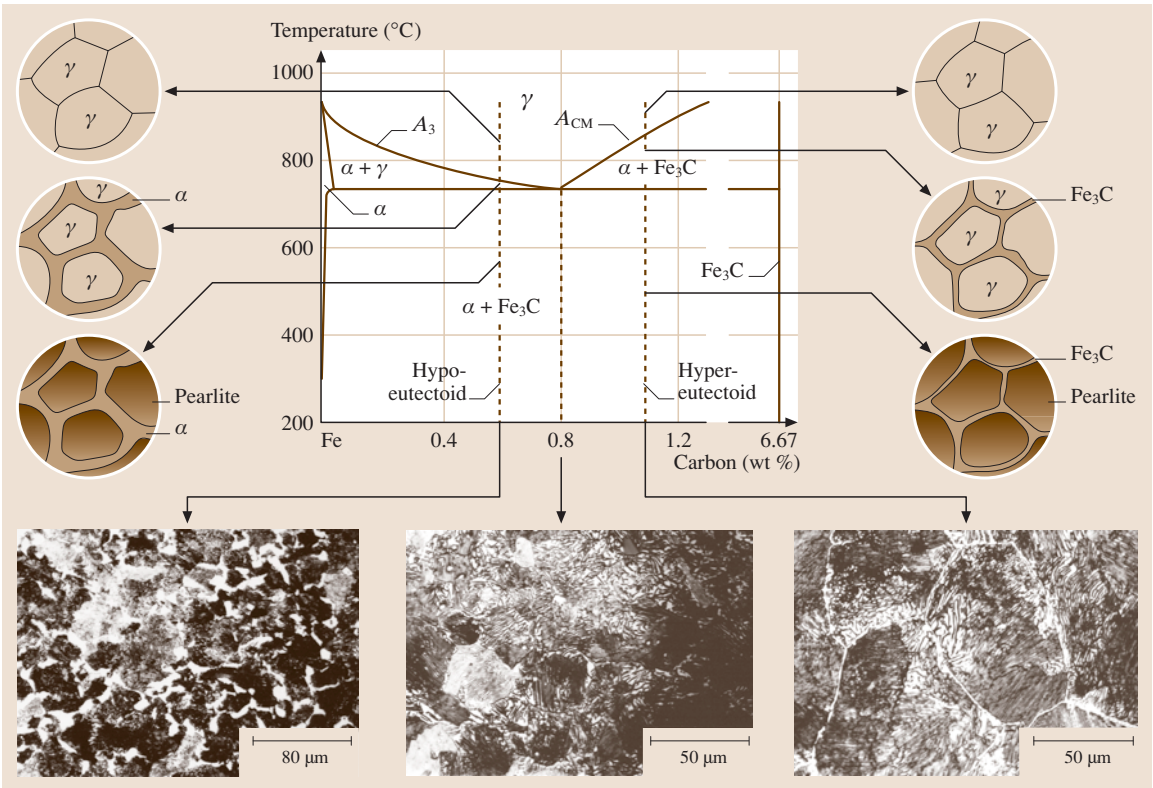


Fig. 3.128 Microstructural evolution of hypo- and hypereutectoid steels upon cooling from the austenitic region of the Fe−C diagram (after [3.54])

equilibrium can easily be derived from microstructure diagrams such as the one shown in Fig. 3.129.

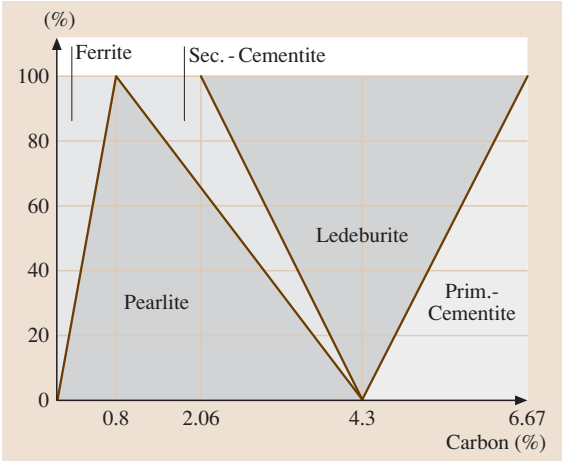


Fig. 3.129 Composition of Fe−C alloys in dependence on the carbon content

Heat Treatments

Since a modification in the atomic configuration requires diffusion (Sect. 3.1.2) of the atoms to occupy the appropriate lattice sites, phase transitions in the solid state are typically time dependent. Therefore, the equilibrium phases of the iron–carbon phase diagram only appear upon slow cooling or after sufficiently long heat treatments. The microstructures shown in Fig. 3.128, on the other hand, arise only under specific cooling conditions. By altering the time–temperature path metastable phases as well as totally different microstructures can be formed. Heat treatment procedures differ with regard to the following parameters:

- Way of heating
- Holding temperature
- Holding time
- Way of cooling (for example, cooling in air, oil, water or furnace)

According to the way of cooling a principle division of the heat treatment procedures is widely used:

1. Annealing treatments (slow cooling close to equilibrium)
2. Hardening treatments (fast cooling)

In the following the main parameters and objectives of frequently used heat treatment procedures for steels are summarized using the above scheme.

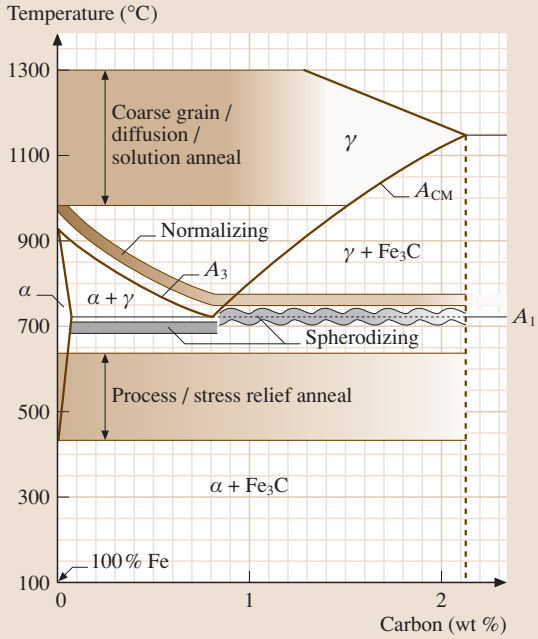
Annealing Treatments. In Fig. 3.130 the temperatures and main parameters of frequently used annealing treatments of steels are shown [3.119].

Normalizing. Due to the heat flow during cooling of castings, in the heat-affected zone of welding joints as wells as after cold or warm rolling, the microstructure of steels can be extremely inhomogeneous. Normalizing, therefore, first of all serves to homogenize the steel and should result in a fine-grained microstructure (grain size $< 100\,\mu m$). To do so, the workpiece is austenitized by heating to temperatures 30–50 K above the A_3 line, i. e., into the region of the γ-Fe solid solution, or above the A_1 line in case of hypereutectoid steels. Subsequent cooling in air leads to complete new formation of fine-grained pearlitic–ferritic or pearlitic–cementitic microstructures (Fig. 3.131a) with high strength and high toughness. The strength and toughness of steels in the normalized condition strongly depend on the carbon content. As shown in Fig. 3.132 the strength reaches its maximum at about 1.0% carbon. Toughness, as characterized by the impact energy of Charpy tests, gives the steel a brittle behavior at carbon concentrations as low as 0.8%.

Spherodizing. The cold workability of normalized steels is normally not sufficient to gain the deformation degree desired. However, since the lamellar arrangement of α-Fe and Fe_3C in pearlite is energetically unstable, on heat treatment slightly below the A_1 transformation temperature (or in the case of hypereutectoid steels by oscillating around A_1 to accelerate spherodizing of the cementite network) the cementite lamellae rearrange to form stable spherical particles (Fig. 3.131b). Dislocation movement, which is restricted in pearlite to the region within the small α-Fe lamellae, is now possible in the whole grain interior and the steel can, therefore, be deformed more severely. After annealing, the workpiece is cooled slowly (furnace) to prevent heat stresses from arising.

Process Annealing. Cold deformation as well as hot deformation to a high degree result in the formation of a dense dislocation network, which hinders dislocation movement and therefore further material deformation. A high dislocation density, on the other hand, stores high levels of energy, which encourages complete new formation of the grain structure (recrystallization) upon annealing at temperatures exceeding $T_p = 0.4T_m$. As a consequence the strength and toughness of the recrystallized state reach levels which are close to the undeformed condition and the material can therefore be further deformed.

Coarse-Grain Annealing. Coarse grains are beneficial when the material is machined by chip-removing meth-



Annealing treatments:

- Normalizing:
  30 – 50 K above A_3
  for 20 – 60 min
  cooling in air
- Spherodizing:
  below/around A_1
  for up to 5 h
  slow cooling in furnace
- Stress relief anneal:
  450 – 650 °C
  for 2 – 4 h
  slow cooling in furnace

Hardening:

- Solution annealing:
  30 – 50 K above A_3
  for 20 – 60 min
- Quenching to room temperature:
  media: oil, water, ice water or salt solutions

Tempering:

- Heat treatment between 100 and 650 °C, outdiffusion of carbon

Fig. 3.130 Annealing treatments of steels

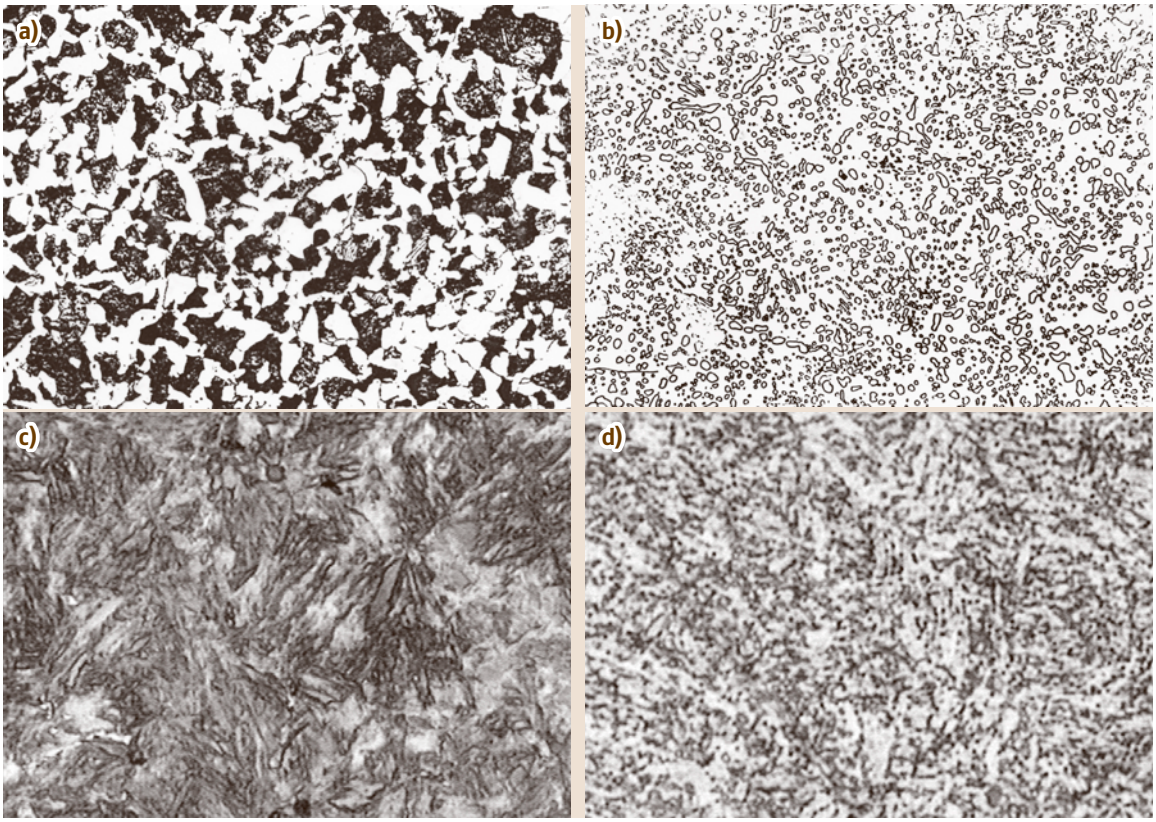


Fig. 3.131a–d Microstructures of steel after different heat treatments: **(a)** C45: normalized, **(b)** C60: spherodized, **(c)** C45: hardened, and **(d)** C45: hardened and tempered at 550 °C

ods because short fragile shear chips are formed. Such a microstructure is the result of a heat treatment at temperatures between 950 and 1200 °C, i. e., in the austenitic region well above the A_3 line. On subsequent cooling in a furnace the coarse-grained γ-Fe solid solution is transformed to a coarse-grained ferritic–pearlitic microstructure. Since the accompanying decrease in toughness deteriorates the steel properties, a final heat treatment (hardening, tempering, etc.) must be made to retransfer the microstructure to a fine-grained state.

Stress-Relief Annealing. Stress-relief annealing serves to relieve stresses in the workpiece which are caused by cold deformation, microstructural transformations, thermal loading or chip-removal working. Stress-relief annealing is usually done at temperatures between 450 and 650 °C for several hours, followed by slow cooling. It does not lead to apparent changes to the microstructure nor does it change the mechanical properties significantly.

Diffusion Annealing. Diffusion annealing is done when segregations (local variations of the chemical composition) have to be compensated and requires temperatures as high as 1000–1300 °C and annealing times as long as 50 h. Since this treatment is very expensive, segregations should be prevented by optimizing the cooling conditions after casting.

Solution Annealing. Solution annealing is predominantly used for austenitic steels and serves to solve (large) precipitates in steels. Annealing at temperatures between 950 and 1200 °C and fast cooling results in a supersaturated solid solution at room temperature. Subsequent aging leads to the formation of small precipitates which lead to a significant strength increase at moderate toughness values.

While the annealing treatments introduced above lead to phase compositions which are close to the equilibrium with increasing cooling rate the transformation behavior of austenite can be completely different.

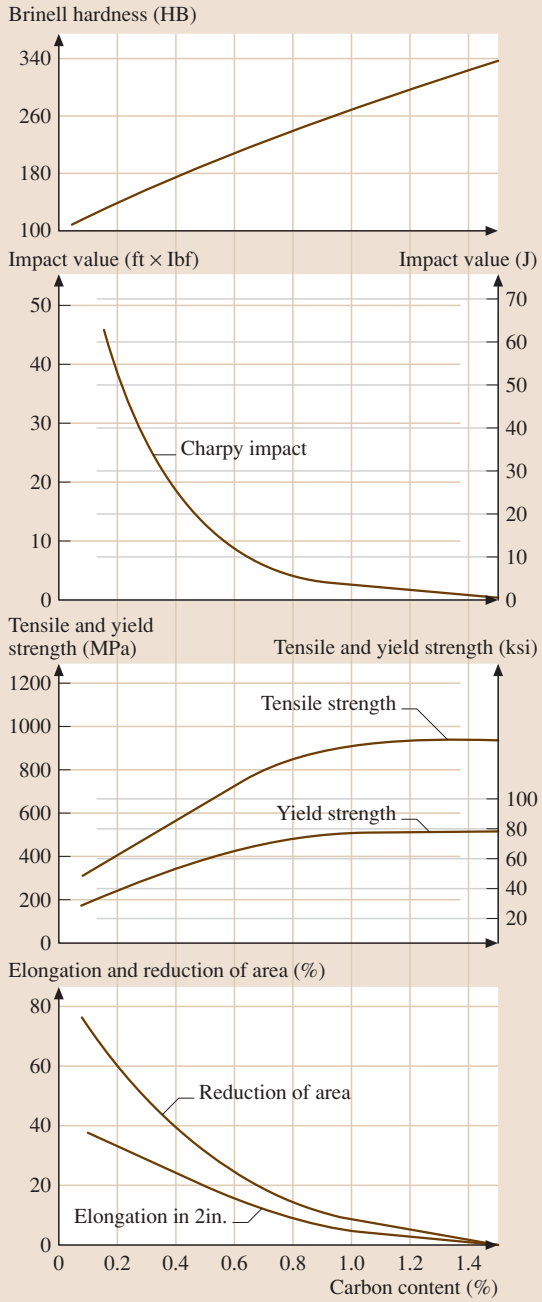


Fig. 3.132 Mechanical properties of Fe−C alloys in dependence on the carbon content (after [3.125])

Since the formation and growth of nuclei are diffusion-controlled processes fast cooling from the γ-Fe phase region may suppress the formation of the temperature

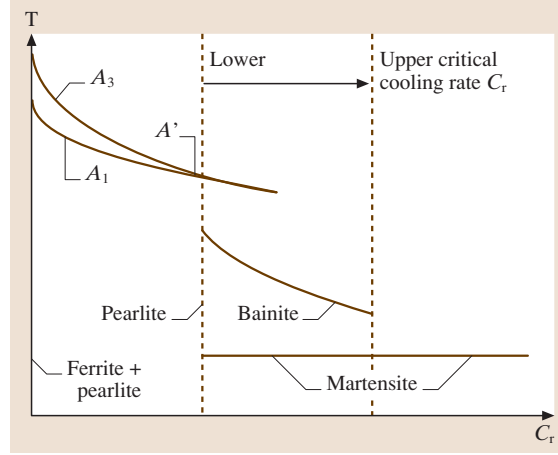


Fig. 3.133 Transition temperatures and products in steels as functions of the cooling rate (C_r)

equivalent equilibrium phases. On increasing cooling rates the transformation temperatures A_1 and A_3 decrease until they coincide at the point of the lower critical cooling rate (Fig. 3.133). At even faster cooling rates *bainite*, which consists of aggregates of plates of ferrite (so-called sheaves), separated by untransformed austenite, *martensite* (see below) or cementite are formed (for a more comprehensive compilation of bainite see [3.124]). Caused by the high cooling rate the diffusivity of Fe is reduced so strongly that the formation of cementite only occurs by diffusion of carbon, while iron transforms from fcc to bcc by a diffusionless shear process. When the cooling rate reaches the upper critical limit the diffusivity of carbon is completely suppressed as well. Due to the supersaturated solution of carbon in α ferrite a distorted body-centered tetragonal (bct) lattice, so-called martensite, is now formed as the product of a phase transition from the fcc γ-lattice. This results in very high strength of the steel but at the expense of low ductility.

The kinetics of the phase transitions from the γ-Fe phase region can be visualized in time–temperature transition (TTT) diagrams, obtained either under isothermal holding conditions or in continuous cooling transition (CCT) diagrams for varying (but fixed) cooling rates. One example for steels is given in Fig. 3.134 showing the CCT diagram of 42MnV7. Depending on the cooling rate, transition of γ-Fe leads to the formation of martensite, bainite or pearlite or a mixture of these at room temperature. The transformation quantity after crossing a transformation region is given in Fig. 3.134 as well as the

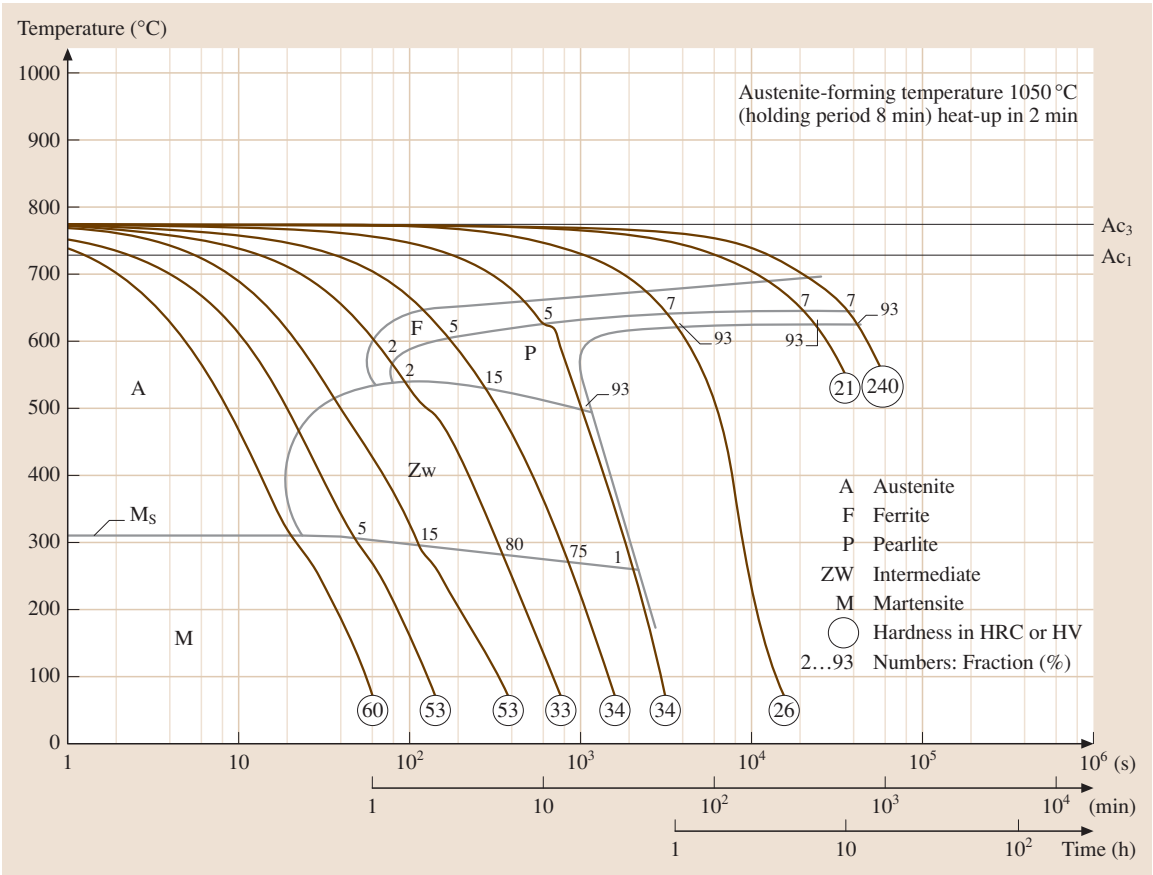


Fig. 3.134 Continuous-cooling-transition (CCT) diagram of 42MnV7

hardness values of the resulting microstructure at room temperature.

Hardening. The basis of hardening is the mechanism of martensitic transformation, which comprises a massive increase in material hardness. In the case of hypoeutectic steels austenitization is done above the A_3 line of the Fe−Fe_3C diagram; in the case of hypereutectic alloys annealing in the two-phase region γ-Fe + Fe_3C is usually sufficient. The distortion caused by the supersaturation of carbon in the α-ferrite increases with increasing carbon content. On the other hand, at least about 0.3% C are necessary to yield a significant increase in strength. The temperature to which the material has to be cooled from the austenite region (without any other nucleus formation to apply during cooling) to form martensite M_S (the *martensite start*) as well as the temperature where the whole microstructure consists of martensite M_f (the *martensite finish*) decrease with increasing carbon content (Fig. 3.135).

Therefore, hardening through the whole cross section of a workpiece with higher carbon content is only possible for low dimensions and the hardening depth can be increased by accelerated cooling or by alloying. The cooling media commonly used are oil, water, ice water or salt solutions. Alloying with Mn, Cr, Mo, and Ni in a concentration range of 1–3% can improve the through hardening capability of the steel. Thermal stresses caused by high temperature gradients superimpose onto transformation-induced stresses and can lead to hardening cracks. The crack sensibility can be reduced by warm-bath hardening, in which a temperature-balancing step above the M_S temperature is done before final quenching to martensite.

Tempering. To gain technically relevant properties, especially suitable toughness values, a final heat treatment

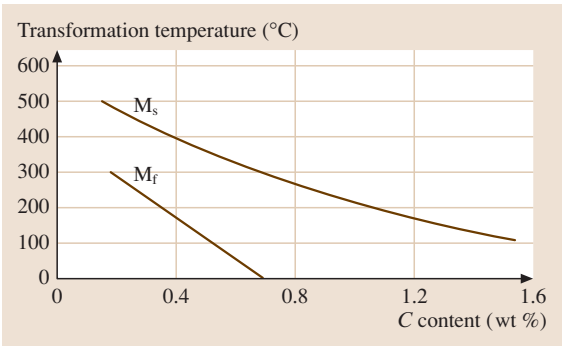


Fig. 3.135 Dependence of the martensite transformation temperatures on the carbon content (after [3.1])

after the hardening process step is required. This tempering step is done below the A_1 line and serves to reduce the brittleness of martensite by means of outdiffusion of carbon from the distorted ferrite lattice. In the temperature range up to about 300 °C, diffusion of carbon leads to a decrease of the lattice distortion which reduces the brittleness but does not lead to a significant change of the strength. This is assured by the formation of small metastable ε-carbides, and the disintegration of residual austenite. Tempering above 300 °C leads to disintegration of the remaining martensite into the formation of ferrite with finely distributed spherical cementite while the ε-carbide transforms to cementite as well. Consequently, the strength is lowered significantly and the toughness increases considerably. Tempering above 450 °C yields a homogeneous, fine-grained microstructure with high toughness and strength, as desired for many structural parts.

In the case of alloyed steels containing Mo, W, and/or V, i. e., tool steels or heat-resistant steels, tempering at temperatures between 450 and 600 °C leads to the formation of small, homogenously distributed carbide precipitates which counteract the strength decrease during annealing (so-called secondary hardening).

Selective Hardening. In many practical applications such as crankshafts, spigots, rolls or gears, high hardness and wear resistance may be required at the surface but at the same time high fracture toughness of the bulk part is required in order not to trade off fatigue strength. Therefore, for these applications, hardening is done only in the near-surface areas of a workpiece. Steels that are suitable for this treatment are plain carbon steels and low-alloy steels with carbon content of 0.3–0.7%. The following treatments are commonly used for direct hardening:

1. Flame hardening, in which the surface of the workpiece is heated to the austenitizing temperature with a gaseous oxygen flame
2. Induction heating, in which a high-frequency coil is used to heat the material surface utilizing the skin effect
3. Beam hardening (electron and laser beam), by which small areas of workpieces can be treated selectively
4. Dip hardening, which is especially suitable for pieces with curved surfaces for which other treatments would be too costly

In the case of steels with less than about 0.25% C the surface has to be enriched in carbon prior to hardening and the workpiece is heated to temperatures between 850 and 950 °C in a carbon-rich atmosphere. It has to be kept in mind that the enriched (up to 0.9% C) surface shows a lower transition temperature then the core with a lower carbon content. Therefore, hardening can be done either from the transition temperature of the core or of the surface. If only the surface region is austenitized, the core is not completely transformed to γ-Fe. This can lead to significant grain growth, and lower toughness values are expected. If, however, the core is fully austenitized, a fine-grained core with significantly higher toughness results after hardening. After surface hardening a tempering treatment at 150–250 °C is usually done.

Nitriding. Nitriding is a thermochemical treatment of steels. The surface area of steels is enriched in nitrogen (usually in an ammonia atmosphere). Nitriding is carried out at relatively low temperatures (495–565 °C) and no quenching is required after the process. Hence, this process leads to relatively little distortion but produces, on the other hand, a relatively shallow case (0.2–0.3 mm). In contrast to carbon-enriched surface layers, nitride layers provide significantly higher temperature resistance (up to 500 °C).

Steel Grades

Effect of Alloying Elements. Besides carbon as the main alloying element, steels generally contain further alloying additions [3.126]. A semantic distinction can be made between:

1. Residual elements, which are not intentionally added to the steel, but result from raw materials and steel-making practices
2. Alloying elements, which are added to cause changes in the properties of steels

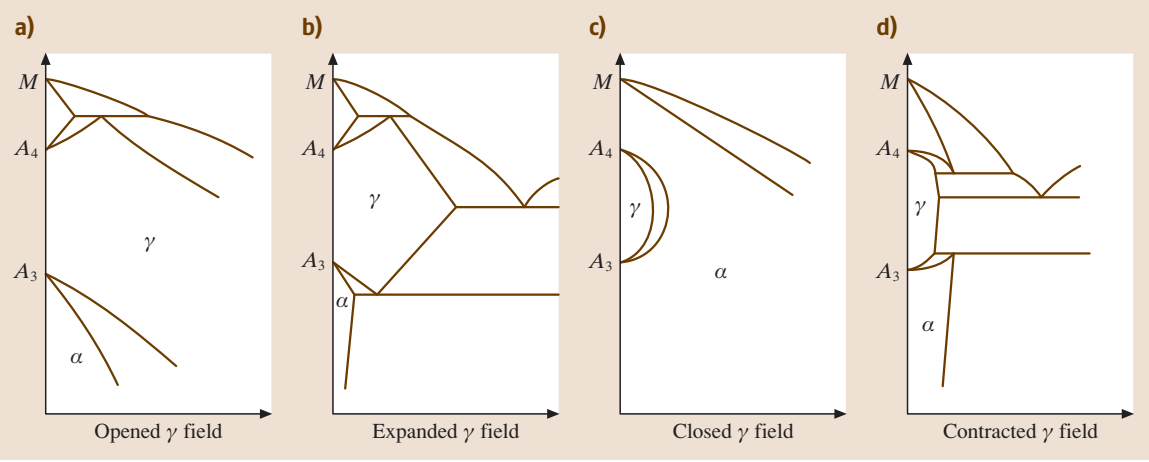


Fig. 3.136a–d Classification of iron alloy phase diagrams: **(a)** opened γ-phase field, **(b)** expanded γ-phase field, **(c)** closed γ-phase field, and **(d)** contracted γ-phase field (after [3.118])

To the residual elements in (1) belong predominantly phosphorus, sulfur, oxygen, hydrogen, manganese, and silicon. Phosphorus, sulfur, oxygen, and hydrogen are usually undesired, because they reduce ductility and toughness. Their content has to be reduced to a harmless level by secondary metallurgy. The only exception is the group of free-machining steels, where sulfur or phosphorus may be added deliberately to improve machinability. The effect of alloying elements of the group (2) can be separated into four different fundamental mechanisms:

a) Change of the phase equilibria in the Fe−C phase diagram
b) Solid solution hardening of elements such as Mn, Si, Ni, Co, Cu, and Al in Fe
c) Formation of carbides
d) Influencing the composition of oxides at surface

Alloying elements can influence the equilibrium diagram in two principal ways:

- By expanding (or opening) the γ-phase field and therefore facilitating the formation of austenite over a wider compositional range. These elements are referred to as γ-stabilizers (Fig. 3.136a,b).
- By contracting (or closing) the γ-phase field, which encourages formation of ferrite over a wider compositional range; these elements are termed α-stabilizers (Fig. 3.136c,d).

Consequently, elements such as Mn and Ni, which open the γ-phase field lead to a decrease of the eutectoid temperature, which facilitates hardening; others such as Ti and Mo lead to an increase of the eutectoid temperature (compare Fig. 3.137). However, most of the alloying elements in steels have in common that they help to decrease the carbon concentration in the eutectoid that provides better hardenability for low-carbon steel grades.

Alloying elements can be separated according to their impact on the iron–carbon phase diagram into four groups [3.118]:

- *Class 1*. These elements open the γ-phase field. To this group belong the most important alloying additions in steels: *manganese* and *nickel*. Cobalt and the inert metals ruthenium, rhodium, palladium, osmium, iridium, and platinum show a similar behavior. With a sufficient amount of nickel or manganese the formation of α-Fe under normal cooling conditions can be suppressed down to room temperature, allowing the formation of austenitic steels. At least 0.3% *manganese* is present in all commercial steel grades. It serves primarily to deoxidize the melt and counteracts the harmful influence of iron sulfide by the formation of manganese sulfide stringers. Excess content of manganese can partly dissolve in the iron lattice, leading to the mentioned solid-solution hardening effect and partly form Mn_3C. Through the opening of the γ-phase field the critical cooling rate is considerably decreased, allowing better hardenability of the steel. With increasing Mn content the amount of C in the steel can be reduced while retaining a constant strength level, which finally leads to improved ductility. The hot working capability is improved

at a Mn content of up to 2% since it reduces the susceptibility to hot shortness. However, if the manganese content is increased above 1.8% the steel tends to become air-hardened with resulting impairment of the ductility. Between 5 and 12% Mn the steel becomes martensitic even after slow cooling. At Mn contents above 12% and high C contents the austenite phase retains down to room temperature. Under impact loading such steels can be strongly cold worked at the surface while the core remains ductile.

In contrast to manganese *nickel* does not form any carbon compounds in steels. Up to a content of about 0.5% it is primarily an efficient ferrite strengthener. This is additionally intensified by a refinement of the pearlite lamellae. As in the case of manganese, with increasing Ni content and hence decreasing transition temperature, hardenability is improved (Fig. 3.138). The sudden drop in Ar_1 temperature at 8–10% nickel encourages the formation of martensite, while above 24% Ni this transformation is depressed below room temperature. As shown in the lower part of Fig. 3.138 the mechanical properties behave accordingly: while steels with more than 10% nickel have a high tensile strength the elongation drops from about 20% at nickel contents below 8% to about 10% in the martensitic region. Above 24% Ni and thus stabilization of austenite at room temperature, the tensile strength decreases and the material becomes ductile, tough, and workable. The effect of increasing Ni and C contents on the microstructure is shown in the *Guillet* diagram for a constant cooling rate in Fig. 3.138. Steels with high nickel content also show a low CTE. The so-called *Invar* alloy, containing 36% nickel, 0.2% carbon, and 0.5% manganese, has a thermal expansion coefficient which is nearly zero over the temperature range 0–100 °C. These alloys are therefore used in clocks, tapes, and wire measurements.

- *Class 2*. These elements expand the γ-phase field. The most important elements belonging to this group are carbon and nitrogen. *Copper*, zinc, and

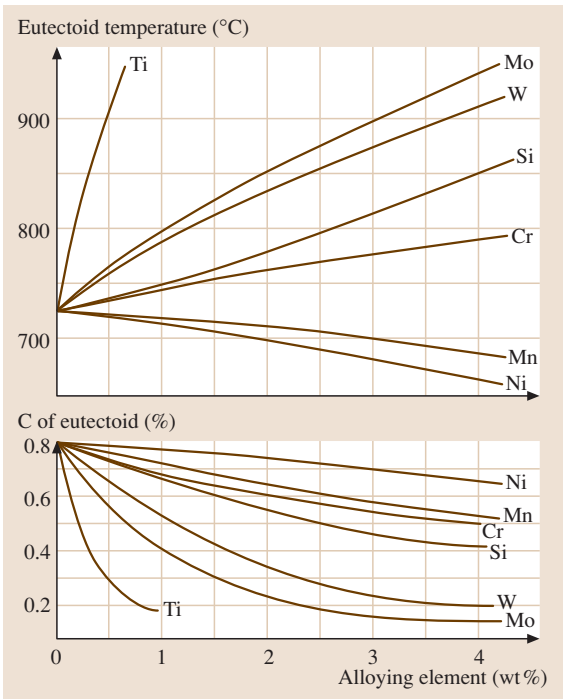


Fig. 3.137 Effect of alloying additions on the eutectoid temperature and the carbon concentration of the eutectoid in steels (after [3.126])

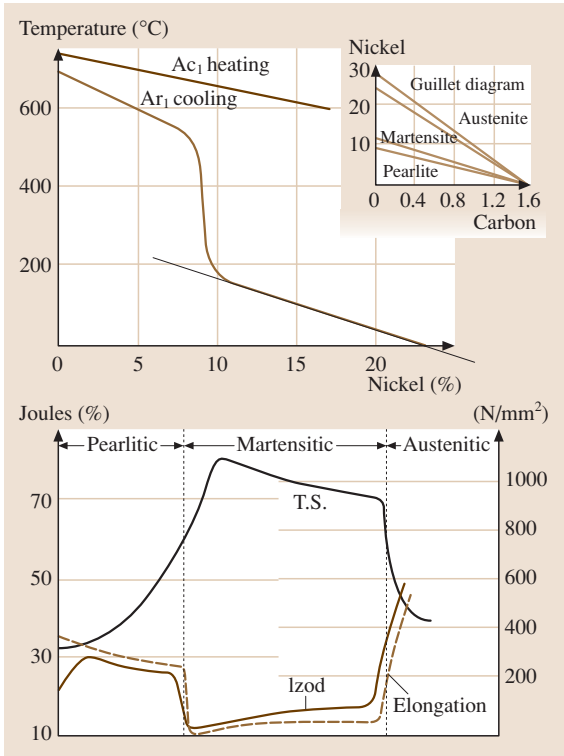


Fig. 3.138 Effect of nickel on transition temperatures and mechanical properties of 0.2% carbon steels cooled at a constant rate (after [3.123])

gold show a similar effect. Cu in amounts exceeding 0.2% is beneficial to atmospheric corrosion resistance for carbon and low-alloy steels. Those steels are referred to as weathering steels and are used in the building industry. Furthermore, copper increases the yield strength and, at a content of more than 0.3%, age hardening is possible. However, copper exaggerates surface defects (grain boundaries), leading to a high surface sensitivity during hot rolling. It is therefore sometimes regarded as a *steel pester*.

- *Class 3*. These elements close the γ-phase field. Elements which restrict the formation of γ-Fe to a small area appearing like a loop include *silicon*, *aluminum*, beryllium, phosphor as well as the carbide-forming elements *titanium*, vanadium, molybdenum, tungsten, and *chromium*. In other words there are more elements which encourage the formation of bcc iron. Note that the normal heat treatment processes which are based on a $\gamma - \alpha$ transformation are no longer available.

When adding *chromium* to steel, most often its capability to increase the resistance to corrosion and oxidation is considered. However, chromium also improves the hardenability by decreasing the critical hardening rate. Furthermore, it raises the high-temperature strength and improves abrasion resistance in high-carbon compositions through carbide formation. Since the carbides are stable at high temperatures the solution anneal temperature has to be increased. In combination with nickel, chromium stabilizes austenite and a steel that superimposes the positive properties of chromium, i. e., high hardness and resistance to wear, and those of nickel, i. e., high strength, ductility, and toughness, is created. The effect of tempering a nickel–chromium steel on the room-temperature mechanical properties is shown in Fig. 3.139. Note that there is a distinct minimum in the Izod impact curve in the temperature range 250–450 °C, known as embattlement. This is caused by the grain boundary enrichment with alloying elements such as Mn and Cr during austenitization, which leads to enhanced segregation of embattling elements such as P, Sn, Sb, and As on slow cooling from 600 °C. This could be prevented by increasing the cooling rate during hardening. However, as shown in Fig. 3.139, Izod (2) addition of *molybdenum* significantly reduces the tempering embrittlement at intermediate temperatures and increases the high-temperature tensile

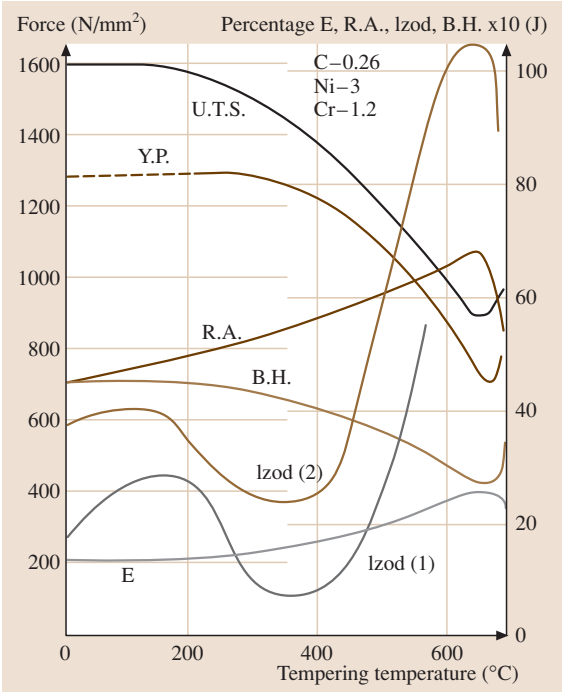


Fig. 3.139 Effect of tempering on the mechanical properties of nickel-chromium steel, C 0.26, Ni 3, Cr 1.2, 29 mm diameter, bars hardened in oil from 830 °C. Izod (2) for steel with 0.25% molybdenum added (U.T.S.: ultimate tensile strength; Y.P.: yield point; E.: elongation; R.A. reduction in area; B.H.: Brinell hardness) (after [3.123])

and creep strength of the steel. Ni–Cr–Mo steels are therefore widely used for ordnance and turbine rotors.

Aluminum, *silicon*, and *titanium* are commonly used as deoxidizers. Furthermore, aluminum and titanium can limit grain growth when added to steel in specific amounts. This is of vital for preventing grain coarsening during solution annealing prior to hardening. Titanium is, however, one of the strongest carbide formers (Fig. 3.140) and, since its

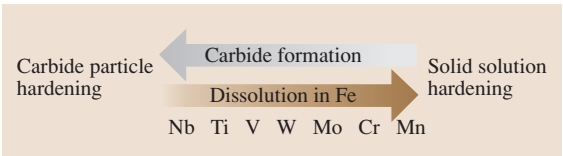


Fig. 3.140 Tendency of alloying elements to form carbides in steels, and vice versa dissolution in Fe lattice

Table 3.15 SAE–AISI system of designation for carbon and alloy steels [3.123]

Nummerals and digits	Type of steel and nominal alloy content (%)
Carbon steels	
10xx [a]	Plain carbon
11xx	Resulfurized
12xx	Resulfurized and rephosphorized
15xx	Plain carbon (max. Mn range 1.00–1.65)
Manganese steels	
13xx	Mn 1.75
Nickel steels	
23xx	Ni 3.50
25xx	Ni 5.00
Nickel–chromium steels	
31xx	Ni 1.25; CR 0.65 and 0.80
32xx	Ni 1.75; Cr 1.07
33xx	Ni 3.50; Cr 1.50 and 1.57
34xx	Ni 3.00; Cr 0.77
Molybdenum steels	
40xx	Mo 0.20 and 0.25
44xx	Mo 0.40 and 0.52
Chromium–molybdenum steels	
41xx	CR 0.50, 0.80 and 0.95; Mo 0.12, 0.20, 0.25 and 0.30
Nickel–chromium–molybdenum steels	
43xx	Ni 1.82; Cr 0.50 and 0.80; Mo 0.25
43BVxx	Ni 1.82; Cr 0.50; Mo 0.12 and 0.25; V 0.03 min
47xx	Ni 1.05; Cr 0.45; Mo 0.20 and 0.35
81xx	Ni 0.30; Cr 0.40; Mo 0.120
86xx	Ni 0.55; Cr 0.50; Mo 0.20
87xx	Ni 0.55; Cr 0.50; Mo 0.25
88xx	Ni 0.55; Cr 0.50; Mo 0.35
93xx	Ni 3.25; Cr 1.20; Mo 0.12
94xx	Ni 0.45; Cr 0.40; Mo 0.12
97xx	Ni 0.55; Cr 0.20; Mo 0.20
98xx	Ni 1.00; Cr 0.80; Mo 0.25
Nickel–molybdenum steels	
46xx	Ni 0.85 and 1.82; Mo 0.20 and 0.25
48xx	Ni 3.50; Mo 0.25
Chromium steels	
50xx	Cr 0.27, 0.40, 0.50 and 0.65
51xx	Cr 0.80, 0.87, 0.92, 0.95, 1.00 and 1.05
50xx	Cr 0.50; C 1.00 min
51xx	Cr 1.02; C 1.00 min
52xx	Cr 1.45; C 1.00 min
Chomium–vanadium steels	
61xx	CR 0.60, 0.80 and 0.95 V 0.10 and 0.15 min
Tungsten–chromium steels	
72xx	W 1.75; Cr 0.75
Silicon–manganese steels	
92xx	Si 1.40 and 2.00; Mn 0.65, 0.82 and 0.85; Cr 0 and 0.65
Boron steels	
xxBxx	B denotes boron steel
Leaded steels	
xxLxx	L denotes leaded steel
Vanadium steels	
xxVxx	V denotes vanadium steel

[a] The xx in the last two digits of these designations indicates that the carbon content (in hundredths of a percent) is to be inserted

carbides are quite stable, they may not dissolve in austenite and can therefore have adverse effects on hardenability. It is used as a stabilizer in corrosion-resistant steels.

- *Class 4*. These elements contract the γ-phase field. This is observed when carbide-forming elements such as tantalum, niobium, and zirconium are present. Boron also belongs to this class of alloying additions. *Zirconium* is primary used in so-called high-strength low-alloy (HSLA) steels to improve their hot-rolling properties.

Classification and Designations. A variety of steel classification systems are in use; they subdivide, for example, with regard to chemical composition, application area, required strength level, microstructure, manufacturing methods, finishing method or the product form (a comprehensive comparison of steels standards is given in [3.127, 128]). Chemical composition is, however, by far the most widely used basis for classification and/or designation of steels. The most commonly used system of designation is those of the American Iron and Steel Institute (AISI) and the Society of Automotive Engineers (SAE), which are based upon a four- or five-digit number, where the first two digits refer to the main alloying elements and the latter two or three digits give the carbon content in wt. %.

Table 3.16 Main groups of UNS designations for iron-based materials

D00001 – D99999	Steels with specified mechanical properties
F00001 – F99999	Cast irons
G00001 – G99999	AISI and SAE carbon and alloy steels (except tool steels)
H00001 – H99999	AISI and SAE H-steels
J00001 – J99999	Cast steels (except tool steels)
K00001 – K99999	Miscellaneous steels and ferrous alloys
S00001 – S99999	Heat- and corrosion-resistant steels (stainless), valve steels, iron-based *superalloys*
T00001 – T99999	Tool steels, wrought and cast

The designation 1020 according SEA–AISI is used, for example, for a carbon steel with nominally 0.2 wt. % C, and the steel 10120 according to SEA–AISI contains 1.2 wt. % C. The various grades of carbon and alloy steels are given in Table 3.15.

The unified numbering system (UNS) for metals and alloys is being used with increasing frequency. It has been developed by ASTM and SAE and other technical societies, trade associations, individual users and producers of metals and alloys, and US government agencies. The system helps to avoid confusion, preventing the use of more than one identification number for the same metal or alloy. Each UNS designation consists of a single-letter prefix followed by five digits. The prefix usually indicates the family class of metals: for example, T for tool steel, S for stainless steel, and F for cast irons, while G is used for carbon and alloy steels. Existing designation systems, such as the AISI–SAE system were incorporated into the UNS system wherever feasible. More information on the UNS system and an in-depth description can be found in SAE J1086 and ASTM E 527. Table 3.16 gives an overview of the main groups of UNS designations for iron-based materials.

The American Society for Testing and Materials (ASTM) standard contains full specifications of specific products, such as A 574 for alloy steel socket-head cap screws, and is oriented towards the performance of the fabricated end product. Theses commonly used steels are not initially included in the SAE–AISI designations.

From a user's viewpoint steels may generally be divided into two main categories, namely *standard steels* and *tool steels*. It is useful to further subdivide *standard steels* according to their chemical composition into three major groups:

1. Carbon steels
2. Alloy steels
3. Stainless steels

Carbon Steels. Carbon steels contain less than 1.65% manganese, 0.6% silicon, and 0.6% copper. According to the SAE standard J142 *General Characteristics and Heat Treatments of Steels* plain carbon steels of the 10xx and 15xx series in Table 3.15 are divided into four groups [3.125]:

- Group I steels with a carbon content of less than 0.15% provide enhanced cold formability and drawability. These steels are therefore used as cold-rolled sheets in automobile panels and appliances and are suitable for welding and brazing. It should however be noted that these alloys are susceptible to grain growth upon annealing after cold working and, as a consequence, exhibit a tendency to embrittlement (strain age-embrittlement).
- Group II steels with carbon contents of 0.15–0.3% show increased strength and hardness and are less suitable for cold forming. The steels are applicable for carburizing or case hardening. As shown above, increasing manganese content supports the hardenability of the core and case, and intermediate manganese levels (0.6–1.0%) are preferential for machining. Carburized plain carbon steels are used for parts which require a hard wear-resistant surface and a *soft* core, for example, small shafts, plungers, and lightly loaded gears.
- Group III steels with medium carbon content of 0.3% to nearly 0.55% can be directly hardened by induction or flame hardening or by cold working. These steels are found in automotive applications and can be used for forgings and for parts which are machined from bar stock.
- Group IV steels with high carbon levels of 0.55% to nearly 1.0% offer improved wear characteristics and high yield strengths and are generally heat treated before use. Since cold-forming methods are not practical for this group of alloys, application parts such as flat stampings and springs are coiled from small-diameter wire. With their good wearing

Table 3.17 Chemical composition and mechanical properties in the as-rolled, normalized, annealed, and quenched-and-tempered condition of some carbon steels [3.125]

SAE –AISI number	Cast or heat chemical ranges and limits (wt.%)				Treatment	Austenitizing/ tempering temperature (°C)	Tensile strength (MPa)	Yield strength (MPa)	Elongation (%)
	C	Mn	P_{max}	S_{max}					
1020	0.17–0.23	0.3–0.6	0.04	0.05	As rolled	–	448.2	330.9	36.0
					Normalized	870	441.3	346.5	35.8
					Annealed	870	394.7	294.8	36.5
1040	0.36–0.44	0.6–0.9	0.04	0.05	As rolled	–	620.5	413.7	25.0
					Normalized	900	589.5	374.0	28.0
					Annealed	790	518.8	353.4	30.2
					Quenched	205	779	593	19
					+ Tempered	650	634	434	29
1095	0.9–1.04	0.3–0.5	0.04	0.05	As rolled	–	965.3	572.3	9.0
					Normalized	900	1013.5	499.9	9.5
					Annealed	790	656.7	379.2	13.0
					Quenched	205	1289	827	10
					+ Tempered	650	896	552	21
1137	0.32–0.39	1.35–1.65	0.04	0.08–0.13	As rolled	–	627.4	379.2	28.0
					Normalized	900	668.8	396.4	22.5
					Annealed	790	584.7	344.7	26.8
					Quenched	205	1082	938	5
					+ Tempered	650	655	483	28

properties typical applications are found in the farm implement industry as plow beams, plow shares, scraper blades, discs, mower knives, and harrow teeth.

The so-called *free-machining grades* are either resulferized (group 11xx steels) or resulferized and rephosphorized carbon steels (group 12xx). These additives enhance their machining characteristics and lower machining costs.

Chemical compositions as well as the mechanical properties of some carbon steels are given in Table 3.17.

Alloy Steels. Alloy steels constitute a category of ferrous metals that exceed the element limits for carbon steels. They contain elements not found in carbon steels such as nickel, molybdenum, chromium (up to 3.99%), cobalt, etc.. The primary function of the alloying elements is to increase the hardenability and to optimize the mechanical properties such as toughness after the final heat treatment. Table 3.18 summarizes the mechanical properties of selected alloy steels in the normalized, annealed, and quenched-and-tempered condition. In the following the alloy steels are divided into five major groups according to their application area [3.125].

Structural steels according to the SAE–AISI system include *carburized steel grades*, *through-hardening grades*, and *nitriding grades*.

Carburizing grades with low alloying combinations such as SAE–AISI 4023 or 4118 have better core properties than plain carbon steels and are hardenable in oil in small cross-sections, resulting in less distortion compared with water-quenched alloys. These alloys are applied as cam shafts, wrist pins, clutch fingers, and other automotive parts. For applications requiring higher core and case hardness such as for automotive gears, universal joints, small hand tools, piston pins, bearings, etc. higher-alloy carburizing steels such as Ni−Mo (SAE–AISI 4620), plain Cr (SAE–AISI 5120) or Ni−Cr−Mo (SAE–AISI 8620) steels are used. Aircraft engine parts, truck transmissions and differentials, rotary rock-bit cutters, and large antifriction bearings are made from high-alloy steels as SAE–AISI 4820 and 9310.

Through-hardening grades in principle contain higher carbon levels than carburized grades. In this group the lower-alloy steels are used for applications

Table 3.18 Mechanical properties of selected alloy steels in the normalized, annealed and quenched-and-tempered condition [3.125]

SAE–AISI number	Treatment	Austenitizing temperature (°C)	Tempering temperature (°C)	Tensile strength (MPa)	Yield strength (MPa)	Elongation (%)
1340	Normalized	870	–	836	558	22
	Annealed	800	–	703	436	26
	Quenched	–	205	1806	1593	11
	+ Tempered	–	650	800	621	22
3140	Normalized	870	–	892	600	20
	Annealed	815	–	690	423	24
	Quenched	–	–	–	–	–
	+ Tempered	–	–	–	–	–
4130 (w)	Normalized	870	–	669	436	26
	Annealed	865	–	560	361	28
	Quenched	–	205	1627	1462	10
	+ Tempered	–	650	814	703	22
4140	Normalized	870	–	1020	655	18
	Annealed	815	–	655	417	26
	Quenched	–	205	1772	1641	8
	+ Tempered	–	650	758	655	22
4150	Normalized	870	–	1155	734	12
	Annealed	815	–	730	379	20
	Quenched	–	205	1931	1724	10
	+ Tempered	–	650	958	841	19
4320	Normalized	895	–	793	464	21
	Annealed	850	–	579	610	29
	Quenched	–	–	–	–	–
	+ Tempered	–	–	–	–	–
4340	Normalized	870	–	1279	862	12
	Annealed	810	–	745	472	22
	Quenched	–	205	1875	1675	10
	+ Tempered	–	650	965	855	19
4620	Normalized	900	–	574	366	29
	Annealed	855	–	512	372	31
	Quenched	–	–	–	–	–
	+ Tempered	–	–	–	–	–
4820	Normalized	860	–	750	485	24
	Annealed	815	–	681	464	22
	Quenched	–	–	–	–	–
	+ Tempered	–	–	–	–	–
5046	Normalized	–	–	–	–	–
	Annealed	–	–	–	–	–
	Quenched	–	205	1744	1407	9
	+ Tempered	–	650	786	655	24

Table 3.18 (cont.)

SAE–AISI number	Treatment	Austenitizing temperature (°C)	Tempering temperature (°C)	Tensile strength (MPa)	Yield strength (MPa)	Elongation (%)
5140	Normalized	870	–	793	472	22.7
	Annealed	830	–	572	293	29
	Quenched	–	205	1793	1641	9
	+ Tempered	–	650	758	662	25
5160	Normalized	855	–	957	531	18
	Annealed	815	–	723	276	17
	Quenched	–	205	2220	1793	4
	+ Tempered	–	650	896	800	20
6150	Normalized	870	–	940	616	22
	Annealed	815	–	667	412	23
	Quenched	–	205	1931	1689	8
	+ Tempered	–	650	945	841	17
8630	Normalized	870	–	650	430	24
	Annealed	845	–	564	372	29.0
	Quenched	–	205	1641	1503	9
	+ Tempered	–	650	772	689	23
8740	Normalized	870	–	929	607	16
	Annealed	815	–	695	416	22
	Quenched	–	205	1999	1655	10
	+ Tempered	–	650	986	903	20
9255	Normalized	900	–	933	579	20
	Annealed	845	–	774	486	22
	Quenched	–	205	2103	2048	1
	+ Tempered	–	650	993	814	20
9310	Normalized	890	–	907	571	19
	Annealed	845	–	820	440	17
	Quenched	–	–	–	–	–
	+ Tempered	–	–	–	–	–

in small sections or in larger sections that may not have optimal properties but allow weight savings due to the higher strength of the alloys. Typical examples are manganese steels (SAE–AISI 1330–45), which are used for high-strength bolts, molybdenum steels (SAE–AISI 4037–4047), and chromium steels (SAE–AISI 5130–50), which are used for automotive steering parts, and low-Ni−Cr−Mo steels (SAE–AISI 8630–50), which are used for small machinery axles and shafts. Heavy aircraft or truck parts or ordnance materials require higher-alloy structural steels, such as SAE–AISI 3430 or 86B45. There are several constructional alloy steels which are used for specialized applications; for example, SAE–AISI 52100 steels are used almost exclusively for ball-bearing applications and the chromium steels SAE–AISI 5150 and 5160 were developed for spring steel applications.

Steels that belong to the *nitriding grades* are in most cases either medium-carbon and chromium-containing low-alloy steels, which are covered by the SAE–AISI (for example, 4100, 4300, 5100, 6100, 8600, 9300, and 9800 group) or Al-containing (up to 1%) low-alloy steels, which are not described by SAE–AISI designations but have simple names such as "Nitralloy". Typical applications for nitride grades include gears designed for low contact stresses, spindles, seal rings, and pins.

Low-carbon quenched-and-tempered steels typically contain less than 0.25% C and less than 5% alloy additions. Economical points of view have driven the

Table 3.19 Characteristics and uses of HSLA steels according to ASTM standards [3.125]

ASTM specification [a]	Title	Alloying elements [b]	Available mill forms	Special characteristics	Intended uses
A 242	High-strength low-alloy structural steel	Cr, Cu, N, Ni, Si, Ti, V, Zr	Plate, bar, and shapes $\leq$ 100 mm in thickness	Atmospheric-corrosion resistance four times of carbon steel	Structural members in welted, bolted or riveted construction
A 572	High-strength low-alloy niobium-vanadium steels of structural quality	Nb, V, N	Plate, bar, and sheet piling $\leq$ 150 mm in thickness	Yield strength of 290 to 450 MPa in six grades	Welded, bolded, or riveted structures, but many bolted or riveted bridges and buildings
A 588	High-strength low-alloy structural steel with 345 MPa minimum yield point $\leq$ 100 mm in thickness	Nb, V, Cr, Ni, Mo, Cu, Si, Ti, Zr	Plate, bar, and shapes $\leq$ 200 mm in thickness	Atmospheric-corrosion resistance four times of carbon steel; nine grades of similar strength	Welded, bolded, or riveted structures, but primarily welded bridges and buildings in which weight savings or added durability is important
A 606	Steel sheet and strip hot-rolled steel and cold-rolled, high-strength low-alloy with improved corrosion resistance	Not specified	Hot-rolled and cold-rolled sheet and strip	Atmospheric-corrosion twice that of carbon steel (type 2) or four times of carbon steel (type 4)	Structural and miscellaneous purposes for which weight savings or added durability is important
A 607	Steel sheet and strip hot-rolled steel and cold-rolled, high-strength low-alloy niobium and/or vanadium	Nb, V, N, Cu	Hot-rolled and cold-rolled sheet and strip	Atmospheric-corrosion twice that of carbon steel, but only when copper content is specified; yield strength of 310 to 485 MPa in six grades	Structural and miscellaneous purposes for which greater strength or weight savings are important
A 618	Hot formed welded and seamless high-strength low-alloy structural tubing	Nb, V, Si, Cu	Square, rectangular round and special-shape structural welded or seamless tubing	Three grades of similar yield strength; may be purchased with atmospheric-corrosion resistance twice that of carbon steel	General structural purposes include welded, bolted or riveted bridges and buildings
A 633	Normalized high-strength low-alloy structural steel	Nb, V, Cr, Ni, Mo, Cu, N, Si	Plate, bar, and shapes $\leq$ 150 mm in thickness	Enhanced notch toughness; yield strenth of 290 to 415 MPa in five grades	Welded, bolted or reveted structures for service at temperatures at or above −45 °C
A 656	High-strength low-alloy, hot rolled structural vanadium-alluminum-nitrogen and titanium-aluminum steels	V, Al, N, Ti, Si	Plate, normally $\leq$ 16 mm in thickness	Yield strength of 552 MPa	Truck frames, brackets, crane booms, mill cars and other applications for which weight savings are important

[a] For grades and mechanical properties

[b] In addition to carbon manganese, phosphorus, and sulfur. A given grade may contain one or more of the listed elements, but not necessarily all of them; for specified compositional limits

[c] Obtained by producing killed steel, made to fine-grain practice, and with microalloying elements such as niobium, vanadium, titanium, and zirconium in the composition

Table 3.19 (cont.)

ASTM specification [a]	Title	Alloying elements [b]	Available mill forms	Special characteristics	Intended uses
A 690	High-strength low-alloy steel H-piles and sheet piling	Ni, Cu, Si	Structural-quality H-pills and sheet piling	Corrosion resistance two to three times greater than that of carbon steel in the splash zone of marine structures	Dock walls sea walls Bulkheads, excavation and similar structures exposed to seawater
A 709, grade 50 and 50 W	Structural steel	V, Nb, N, Cr, Ni, Mo	All structural shape groups and plate $\leq$ 100 mm thickness	Minimum yield strength of 345 MPa, grade 50 W is a weathering steel	Bridges
A 714	High-strength low-alloy welded and seamless steel pipe	V, Ni, Cr, Mo, Cu, Nb	Pipe with nominal pipe-size diameters of 13 to 660 mm	Minimum yield strength of $\leq$ 345 MPa and corrosion resistance two or four times that of carbon steel	Piping
A 715	Steel sheet and strip hot-rolled, high-strength low alloy with improved formability	Nb, V, Cr, Mo, N, Ti, Zr, B	Hot-rolled sheet and strip	Improved formability [c] compared to a A 606 and A 607; yield strength of 345 to 550 MPa in four grades	Structural and miscellaneous applications for which high strength, weight savings, improved formability and good weldability are important
A 808	High-strength low-alloy steel with improved notch toughness	V, Nb	Hot-rolled plate $\leq$ 65 mm in thickness	Charpy V-noth impact energies of 40–60 J (40–60 ft lfb) at −45 °C	Railway tank cars
A 812	High-strength low-alloy steel	V, Nb	Steel sheet in coil form	Yield strength of 450–550 MPa	Welded layered pressure vessels
A 841	Plate produced by thermomechanical controlled processes	V, Nb, Cr, Mo, Ni	Plates $\leq$ 100 mm in thickness	Yield strength of 310–345 MPa	Welded pressure vessels
A 847	Cold-formed, welded and seamless high-strength low-alloy structural rubbing with improved atmospheric corrosion resistance	Cu, Cr, Ni, Si, V, Ti, Zr, Nb	Welded rubbing with maximum periphery of 1625 mm and wall thickness of 16 mm or seamless tubing with maximum periphery of 810 mm and wall thickness of 13 mm	Minimum yield strength $\leq$ 345 MPa with atmospheric-corrosion twice that of carbon steel	Round, square, or specially shaped structural tubing for welded, riveted or bolted construction of bridges and buildings
A 860	High-strength butt-welding fittings of wrought high-strength low-alloy steel	Cu, Cr, Ni, Mo, V, Nb, Ti	Normalized or quenched-and-tempered wrought fittings	Minimum yield strength $\leq$ 485 MPa	High-pressure gas and oil transmission lines
A 871	High-strength low-alloy steel with atmospheric corrosion resistance	V, Nb, Ti, Cu, Mo, Cr	As-rolled plate $\leq$ 35 mm thickness	Atmospheric-corrosion resistance four times that of carbon structural steel	Tubular structures and poles

[a] For grades and mechanical properties

[b] In addition to carbon manganese, phosphorus, and sulfur. A given grade may contain one or more of the listed elements, but not necessarily all of them; for specified compositional limits

[c] Obtained by producing killed steel, made to fine-grain practice, and with microalloying elements such as niobium, vanadium, titanium, and zirconium in the composition

development of these steels and the choice of alloying additions accordingly. With their low carbon content these steels have high ductility and notch toughness and are suitable for welding while still offering high

Table 3.20 Compositions and properties of some widely used stainless steels [3.129]

AISI–SAE grade	Nominal composition (wt.%) C	Cr	Ni	Others	Condition	Yield strength (MPa)	Tensile strength (MPa)	Elongation (%)
Austenitic grades								
201	0.15	17	5	6.5%Mn	Annealed	310	650	40
304	0.08	19	10		Annealed	205	520	30
					Cold-worked	965	1275	9
304L	0.03	19	10		Annealed	205	520	30
316	0.08	17	12	2.5%Mo	Annealed	205	520	30
321	0.08	18	10	0.4%Ti	Annealed	240	585	55
347	0.08	18	11	0.8%Nb	Annealed	240	620	50
Ferritic grades								
430	0.12	17			Annealed	205	450	22
442	0.12	20			Annealed	275	520	20
Martensitic grades								
416	0.15	13		0.6%Mo	Quenched and tempered	965	1240	18
431	0.2	16	2		Quenched and tempered	1035	1380	16
440C	1.1	17		0.7%Mo	Quenched and tempered	1895	1965	2
Nonstandard (precipitation-hardened) grades								
17–4	0.07	17	4	0.4%Nb	Age-hardened	1170	1310	10
17–7	0.09	17	7	1.0%Al	Age-hardened	1585	1650	6

yield strengths (approximately 340–900 MPa). In addition, they have two to six times higher corrosion resistance than that of plain carbon steels. Depending on the final treatment these steels could be either martensitic, bainitic, and, in some compositions, ferritic. These steels are not covered by SAE–AISI designations but most of them can, however, be find in ASTM specifications such as A514, A517, and A543. Thanks to the high strength and toughness values these steels can be applied at lower final costs than plain carbon steels, which leads to a wide variety of applications. They are used as major members of large steel constructions, pressure valves, earth-moving, and mining equipment.

Ultrahigh-strength steels are a group of alloy steels with yield strengths in excess of 1300 MPa; some have plain-strain fracture toughness levels exceeding 100 MPa$\sqrt{\text{m}}$. Some of these steels are included in the SAE–AISI designation system and have medium carbon contents with low-alloy additions. Examples are steels in the SAE–AISI 4130 series, the higher-strength 4140, and the deeper hardening higher-strength 4340 steels. Starting form the 4340 alloy series numerous modifications have been developed. Addition of silicon, for example, reduces the sensitivity to embrittlement on tempering at low temperatures (required to keep high strength levels). Addition of vanadium leads to grain refinement, which improves the strength and toughness of the material. Medium-carbon alloys can be welded in the annealed or normalized condition, requiring a further heat treatment to retrieve the desired strength. If high fracture toughness as well as high strength is specifically desired, as for aircraft structural components, pressure vessels, rotor shafts for metal-forming equipment, drop hammer rods, and high-strength shock-absorbing automotive parts, high nickel (7–10.5%) and Co (4.25–14.50%) contents are used as primary alloying elements. While offering a plane-strain fracture toughness of 100 MPa$\sqrt{\text{m}}$ the HP-9-4-30 steel can have a tensile strength as high as 1650 MPa. Furthermore, the steel can be hardened to martensite in sections up to 150 mm thick. The AF 1410 steel (developed by the US Air Force) has an ultimate tensile strength (UTS) of 1615 MPa and a K_{IC} value of 154 MPa$\sqrt{\text{m}}$.

The group of *alloy steels for elevated- or low-temperature applications* includes two different alloying systems. For high-temperature applications chromium–molybdenum steels offer a good combination of oxidation and corrosion resistance (provided by

the chromium content of up to 9%) on the one hand and high strength at elevated temperatures (provided by the molybdenum content of 0.5–1.0%) on the other. These steels can be applied at temperatures up to 650 °C for pressure vessels and piping in the oil and gas industries and in fossil-fuel and nuclear power plants. In low-temperature service applications such as storage tanks for liquid hydrocarbon gases and structures and machinery design for use in cold regions, ferritic steels with high nickel content (approximately 2–9%) are typically used.

Another important category of alloy steels are the *high-strength low-alloy steels* (HSLA). HSLA steels, or microalloyed steels, are designed to meet specific mechanical properties rather than a chemical composition. So the chemical composition can vary for different end-product thicknesses with still retaining specific mechanical properties. The low carbon content of these steels (0.05–0.25%) allows good formability and excellent weldability. Further alloying elements are added to meet the application requirements (Table 3.19), resulting in a division into six categories, as follows:

- Weathering steels, where small amounts of copper and phosphorous are added to improve atmospheric corrosion resistance
- Microalloyed ferritic–pearlitic steels, with small amounts (less than 0.1%) of carbide-forming elements such as niobium, vanadium or titanium which enable precipitation strengthening and grain refinement
- As-rolled pearlitic steels, with high strength, toughness, formability, and weldability, which have carbon, manganese, and further additions
- Acicular ferrite (low-carbon bainite) steels (less than 0.08% C), which offer an excellent combination of high yield strength, weldability, formability, and good toughness
- Dual-phase steels, with martensitic portions finely dispersed in a ferritic matrix. These steels have high tensile strength and sufficient toughness
- Inclusion-shape-controlled steels, in which the shape of sulfide is changed from elongated stringers to small, dispersed, near-spherical globules to improve ductility and toughness; elements which are suitable are, e.g., Ca, Zr, and Ti

The allocation to a specific group is not rigorous; many of these steels have properties which would also allow allocation to other groups mentioned.

Stainless steels. Stainless steels in general contain at least 12% chromium, which forms a thin protection layer at the surface (Cr−Fe−oxide) when exposed to air [3.129]. As shown above, chromium stabilizes the ferrite to remain stable up to the melting point, presuming, however, a low carbon content. Stainless steels can be differentiated depending on their crystal structure or the acting strengthening mechanisms according to Table 3.20. Ferritic stainless steels are relatively inexpensive and contain as much as 30% chromium with typically less than 0.12% C. They show good strength and intermediate ductility. Martensitic stainless steels typically contain less than 17% chromium to contract the austenitic region not too strongly but have a higher C content of up to 1.0%. These alloys are used for high-quality knifes, ball bearings or fittings.

Austenitic stainless steels are formed by the addition of nickel, offer high ductility, and are intrinsically not ferromagnetic. These alloys are well suited for high-temperature applications because of their high creep resistance and, thanks to their high toughness at low temperatures, for cryogenic service as well.

Precipitation-strengthened stainless steels contain additions such as Al, Nb or Ta, which form precipitates such as Ni_3Al during heat treatment and can have very high strength levels.

Stainless steels with duplex microstructure consist of about 50% ferrite and austenite each. They show an ideal combination of strength, toughness, corrosion resistance, formability, and weldability, which no other stainless steel can supply.

Tool steels. *Tool steels* are made to meet special quality requirements, primarily due to their use in manufacturing processes as well as for machining metals, woods, and plastics [3.130]. Some examples are cutting tools, dies for casting or forming, and gages for dimensional tolerance measurements. Tool steels are very clean ingot-cast wrought products with medium (minimum 0.35%) to high carbon content and high alloy (up to 25%) contents, making them extremely expensive. They must withstand temperatures up to 600 °C and should in addition have the following properties:

- Generally a high hardness to resist deformation.
- Resistance to wear for economical tool life, which depends directly on hardness; this can be increased by alloying with carbide-forming elements such as W and Cr.
- Dimensional stability. Dimensional changes of tools can be caused by microstructural alteration, by

Table 3.21 Chemical composition and usage of selected tool steels [3.129]

Designation		Composition in % (with emphasis to show differences between steels belonging to each group)								Typical uses
AISI-SAE	UNS no.	C	Mn	Si	Cr	V	W	Mo	Co	
Water-hardening grades										
W1	T72301	0.7–1.5				0.1 max				Cold-heading dies, woodworking tools, etc.
W2	T72302	0.85–1.5				0.15–0.35				
Shock-resisting tool steels										
S1	T41901	0.4–0.55	0.1–0.4	0.15–1.2	1.00–1.80		1.50–3.00	0.5 max		Chisels, hammers, rivet sets, etc.
S2	T41902	0.4–0.55	0.3–0.5	0.9–1.2				0.30–0.60		
Oil-hardening cold-work tool steels										
O1	T31501	0.85–1.00	1.0–1.4		0.4–0.6		0.4–0.6			Short-run cold-forming dies, cutting tools
O2	T31502	0.85–0.95	1.4–1.8		0.5 max					
Air-hardening medium-alloy cold-work tool steels										
A2	T30102	0.95–1.05	1.00 max		4.75–5.5		0.9–1.4			Thread rolling and slitting dies, intricate die shapes
A6	T30106	0.65–0.75	1.8–2.5		0.9–1.2		0.9–1.4			
High carbon high-chromium cold-work steels										
D2	T30402	1.40–1.60			11–13	1.1 max		0.7–1.2		Uses under 482 °C, gages, long-run forming and blanking dies
D3	T30403	2.00–2.35			11–13	1.0 max				
D4	T30404	2.05–2.40			11–13	1.0 max		0.7–1.2		
Chromium hot-work steels										
H12	T20812	0.30–0.45		0.8–1.2	4.75–5.5	0.5 max	1.0–1.7	1.25–1.75		Al or Mg extrusion dies, die-casting dies, mandrels, hot shears, forging dies
H13	T20813	0.32–0.45		0.8–1.2	4.75–5.5	0.8–1.2		1.1–1.75		
H19	T20819	0.32–0.45		0.2–0.5	4.0–4.75	1.75–2.2	4.0–5.25	0.3–0.55	4.0–4.5	
Tungsten high-speed steels										
H21	T20821	0.26–0.36		0.15–0.5	3.0–3.75	0.3–0.6	8.5–10.0			Hot extrusion dies for brass, nickel, and steel, hot-forging dies
H23	T20823	0.25–0.45		0.15–0.4	11.0–12.75	0.75–1.25	11–12.75			
Tungsten high-speed steels										
T1	T12001	0.65-0.8			3.75–4.5	0.9–1.3	17.25–18.75			Original high-speed cutting steel, most wear-resistant grade
T15	T12015	1.5–1.6			3.75–5	4.5–5.25	11.75–13.0	1.0 max	4.75–5.25	
Molybdenum high-speed steels										
M1	T11301	0.78–0.88			3.5–4.0	1–1.35	1.4–2.1	8.2–9.2		Lower cost than T-type tools
M2	T11302	0.78–0.88			3.75–4.5	1.75–2.2	5.5–6.75	4.5–5.5		
M3	T11313	1.0–1.1			3.75–4.5	2.25–2.75	5.5–6.75	4.75–6.5		
M10	T11310	0.84–0.94			3.75–4.5	1.8–2.2		7.75–8.5		

exceeding the elastic limit, by transformation of the remaining austenite, and by changes of the grain dimensions. The highest dimensional stability is exhibited by ledeburitic cut steels with 12% chromium, which are used for precise cutting and punch tools.

- Working capability. In order to guarantee fault-free operation tools must have a certain working capability, meaning that they should be able to collect elastic distortion energy to a certain degree. Therefore, steels with high toughness and at the same time high yield strength are required.
- Through hardening. The through hardening capability can be improved by alloying with carbide formers which additionally increase wear resistance: Cr, Mo, and Mn.

Tool steels may be categorized into five principal groups; compositions and application examples are given in Table 3.21:

1. Cold-work tool steels, which include the oil-hardening O alloys, the water-hardening W alloys, the high-chromium class D (stainless steel), and the medium-alloy air-hardening class A alloys. Water-hardening grades have high resistance to surface wear but are not suited for high-temperature applications.
2. Shock-resistant tool steels in the S group of alloys, which are the toughest of the tool steels due to the presence of only 0.5% C and low-alloying additions.
3. Hot-work tool steels are class H alloys and include chromium, tungsten, and molybdenum alloys.
4. High-speed steels are either tungsten (T class) or molybdenum (M class) alloys. They have a high hardness of 62–67 HRC and maintain this hardness at service temperatures as high as 550 °C.
5. Special-purpose tool steels are low-alloy (L) or mold tool steels (P).

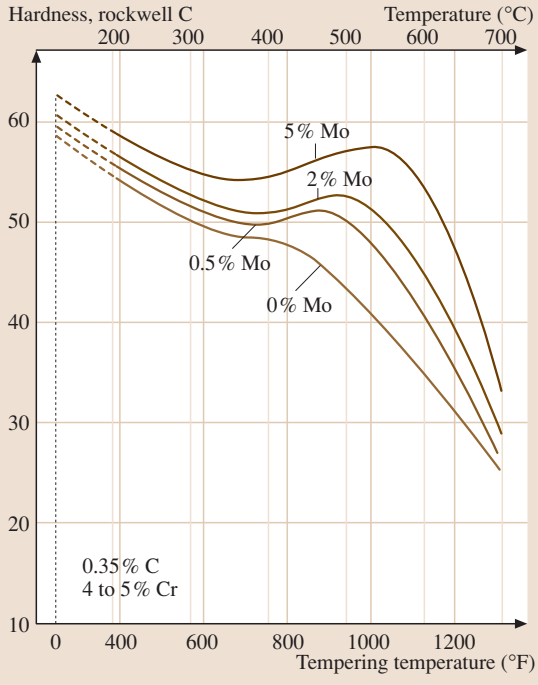


Fig. 3.141 Influence of molybdenum content on the occurrence of a secondary hardening maximum during tempering (after [3.131])

From the compositions given in Table 3.21 it is obvious that the main alloying elements in tool steels are:

1. Chromium to increase hardenability and, if alloyed in excess, forms $Cr_{23}C_6$ for high wear resistance
2. Molybdenum and tungsten, which are strong hard carbide formers $(Mo\text{-}W)_6C$ that can be dissolved in austenite, and precipitate as fine particles in martensite upon annealing (secondary hardening, see Fig. 3.141). Furthermore, they resist growth at low red temperatures.
3. Vanadium, which forms the hardest carbide V_4C_3, and which resists solution in austenite and remains unchanged through heat-treatment cycles.

Steel Castings and Cast Iron

Steel castings and cast iron are preferable for the manufacture of complex geometries at relatively low costs because expensive reworking steps are not necessary or only a few process steps are required to reach the final product [3.132]. Against this advantage there are two important restrictions, namely:

1. The appearance of cast defects and, in conjunction
2. Inferior mechanical properties compared with components prepared by deformation

Iron-carbon cast materials can be roughly divided according to their carbon content into:

1. Steel castings ($C < 2\%$)
2. Cast iron ($C > 2\%$)

Steel castings have a better combination of high strength and ductility compared with cast iron. However, because of their high melting point and strong shrinkage (about 2%) upon cooling the castability of steel castings is poor and the affinity of forming cavities is more pro-

nounced. Therefore, steel castings are only used when high strength and toughness are a must, as in the case of permanent magnet castings and manganese hard castings. A far more broad application spectrum exists for cast irons, which are concentrated on in the following. Cast iron is a very cost-efficient constructional material. The precipitation of carbon (as graphite) during solidification counteracts the normal shrinkage of the solidified metal.

Classification of Cast Iron. Cast irons solidify by the eutectic reaction and are generally ternary alloys of Fe with 2–4% C and 0.5–3% Si. With increasing contents of C and Si and decreasing cooling rate the formation of the stable graphite instead of the unstable cementite is favored. Furthermore, high carbon content and silicon give cast irons excellent castability with melting points appreciably lower than those of steel. Pattern-making is no longer a necessary step in manufacturing cast-iron parts. Many gray, ductile, and alloy-iron components can be machined directly from bars that are continuously cast to near-net shape. Not only does this *parts without patterns* method save the time and expense of pattern-making, but continuous-cast iron also provides a uniformly dense, fine-grained structure, essentially free from porosity, sand, or other inclusions. Cast irons are usually not classified according to their chemical composition. The microstructure of the final product depends strongly upon foundry practice and the shape and size of the castings, which influence the cooling rate; so several entirely different types of iron may be produced starting with the same nominal composition. Thus, cast irons are usually specified by their mechanical properties. A principal classification can be made concerning their microstructure, which depends on the casting conditions:

1. Gray cast iron
2. Ductile cast iron
3. White cast iron
4. Compacted graphite iron
5. Malleable cast iron
6. High-alloy cast irons

Gray cast iron is formed when excess carbon graphitizes during solidification to form separate graphite flakes. The resulting microstructure depends on the cooling rate from the eutectoid temperature downwards (region II in Fig. 3.142). If cooling is fast, pearlite (α-Fe + Fe_3C) is formed from the γ-Fe. If the cooling rate is slow ferrite is formed during transformation and

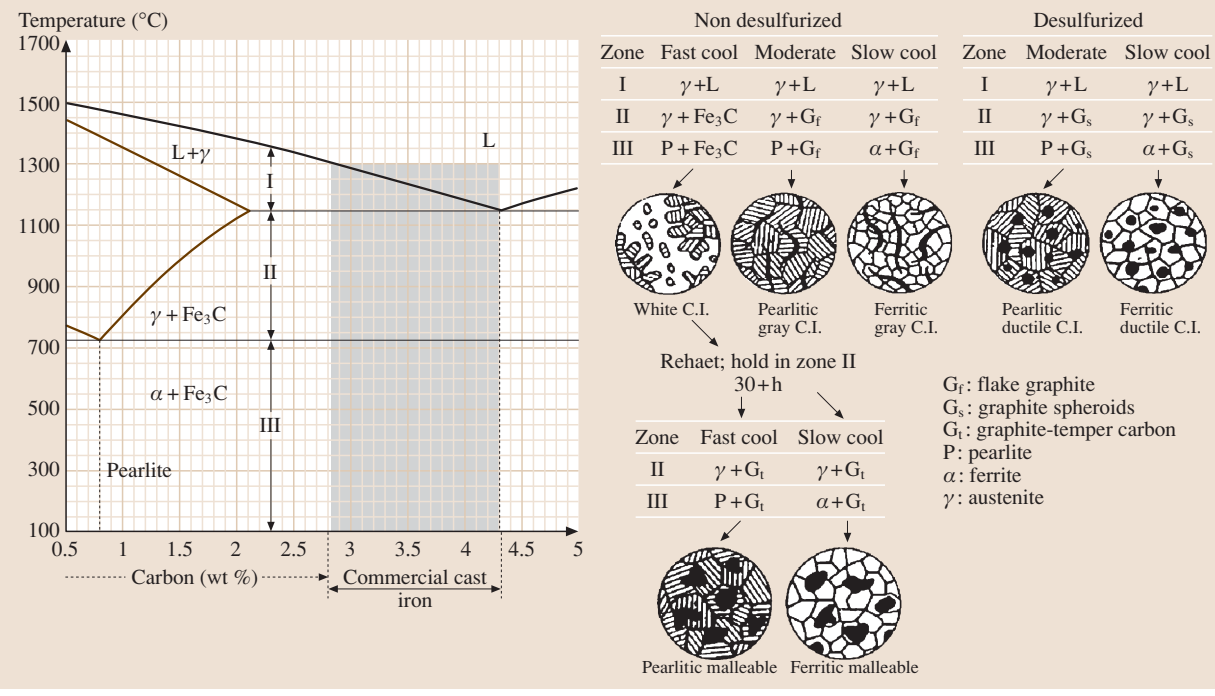


Fig. 3.142 The influence of the casting conditions on the resulting microstructure of cast irons (after [3.133])

Table 3.22 Mechanical properties of forged steel, pearlitic ductile iron, and ADI [3.134]

Mechanical property	Material		
	Forged steel	Pearlitic ductile iron	Grade 150/100/7 ADI
Tensile strength (MPa)	790	690	1100
Yield strength (MPa)	520	480	830
Elongation (%)	10	3	10
Brinnel hardness	262	262	286
Impact strength (ft-lb) (J)	130	40	120

the material has a lower strength compared with the pearlitic gray cast iron. Depending on the cooling rate a mixture of ferrite (surrounding the graphite flakes) and pearlite may be formed as well. The flake-type shape of the graphite in gray cast iron leads to generally brittle behavior. Furthermore, the impact strength of gray cast iron is low and it does not have a distinct yield point. On the other hand, excellent damping against vibrations, excellent wear resistance, and acceptable fatigue resistance are desirable properties of gray cast iron. Typical applications are engine blocks, gears, flywheels, brake discs and drums, and machine bases.

In *ductile iron* the form of the graphite is nodular or spheroidal instead of flake type. This is achieved by the addition of trace amounts of Mg and/or Ce which react with sulfur and oxygen. However, in ductile iron the impurity level has to be controlled more precisely than in gray cast iron since it affects nodule formation. Ductile cast iron exhibits improved stiffness and shock resistance. It has good machinability and fatigue strength as well as high modulus of elasticity, yield strength, wear resistance, and ductility. Damping capacity and thermal conductivity are lower than in gray iron. By weight, ductile gray iron castings are more expensive than gray iron. Ductile iron is used in applications such as valve and pump bodies, crankshafts, in heavy-duty gears or automobile door hinges, and nowadays with increasing frequency also as engine blocks.

Austempered ductile cast iron (ADI) is a subgroup of the ductile iron family but could be treated as a separate class of engineering materials. In contrast to the former, the matrix of this spheroidal graphite cast iron is bainitic (not pearlitic). This microstructure is obtained by isothermal transformation of austenite at temperatures below that at which pearlite forms. In terms of properties, the bainitic matrix has almost twice the strength of pearlitic ductile iron while retaining high elongation and toughness. While exhibiting superior wear resistance and fatigue strength the castability of ADI is not very different from that of other ductile irons, but heat treatment is a critical issue to fully exploit its beneficial properties. For example, the yield strength of ADI is more than three times that of the best cast or forged aluminum. In addition ADI castings weigh only 2.4 times more than Al alloys and are 2.3 times stiffer. ADI is also 10% less dense than steel. Furthermore, for a typical component, ADI costs 20% less per unit weight than steel and half that of Al. A comparison of forged steel, pearlitic ductile iron, and ADI is shown in Table 3.22.

White cast irons are formed trough fast cooling and consist of Fe_3C and pearlite. The origin of this designation is the white-appearing crystalline fracture surface. While having an excellent wear resistance and high compressive strength the principal disadvantage of white cast iron is its catastrophic brittleness. Therefore in most applications white cast iron is only formed on the surface of cast parts, while the core consists of either grey cast iron or ductile iron. Examples of the application of white cast iron are mill liners and shot-blasting nozzles as well as railroad brake shoes, rolling-mill rolls, and clay-mixing and brick-making equipment, crushers, and pulverizers.

Compacted graphite iron (CGI), also known as vermicular iron, can be considered as an intermediate between gray and ductile iron, and possesses many of the favorable properties of each. CGI is difficult to produce successfully on a commercial scale because the alloy additions must be kept within very tight limits. The advantages of CGI compared with gray cast iron are its higher fatigue resistance and ductility, which are at the same level as those of ductile iron. Machinability, however, is superior to that of ductile iron and its damping capacity is almost as good as that of gray iron. This combination and the high thermal conductivity of CGI suggest applications in engine blocks, brake drums, and exhaust manifolds of vehicles.

Malleable iron is white iron that has been converted by a two-stage heat treatment to a condition in which most of its carbon content is in the form of irregularly shaped nodules of graphite, called temper carbon. In contrast to white iron it is malleable

and easily machined. The matrix of malleable cast iron could be ferritic, pearlitic, or martensitic. Ferritic grades are more machinable and ductile, whereas the pearlitic grades are stronger and harder. Malleable-iron castings are often used for heavy-duty bearing surfaces in automobiles, trucks, railroad rolling stock, and farm and construction machinery. The applications are, however, limited to relatively thin-sectioned castings because of the high shrinkage rate and the need for rapid cooling to produce white iron.

High-alloy irons are ductile, gray, or white irons that contain 3% to more than 30% alloy content. Properties achieved by specialized foundries are significantly different from those of unalloyed irons. These irons are usually specified by chemical composition as well as by various mechanical properties. White high-alloy irons containing nickel and chromium develop a microstructure with a martensitic matrix around primary chromium carbides. This structure provides high hardness with extreme wear and abrasion resistance. High-chromium irons (typically, about 16%) combine wear and oxidation resistance with toughness. Irons containing 14–24% nickel are austenitic; they provide excellent corrosion resistance for nonmagnetic applications. The 35% nickel irons have an extremely low coefficient of thermal expansion and are also nonmagnetic and corrosion resistant.

3.7.2 Aluminum and Its Alloys

General Properties

Despite the ten times higher costs for the preparation of primary aluminum compared with that of pig iron, aluminum-based materials are today the second most widely used metallic materials. For structural applications in mechanical engineering their most important advantage is an excellent combination of intrinsically good corrosion and oxidation resistance and high specific strength (strength-to-density ratio; compare Table 3.13) when compared with stainless steels. However, with strength values as low as 45 MPa (1199-O), technically pure aluminum is very soft and requires the addition of alloying elements for most applications. In doing so the strength can be increased manyfold to almost 700 MPa (alloy 7055-T77). Because the specific stiffness (stiffness-to-density ratio) of aluminum alloys ($E \approx 70$ GPa) is comparable to that of steels ($E \approx 210$ GPa), components must have significantly larger dimensions (volume) in order to achieve a stiffness equal to that of their steel counterparts. However, various constructive arrangements, such as tabular or box-shaped hollow sections, locally strengthened rips, flanges welts, and coils, are known and can be used as an alternative to larger components. In contrast to steels Al does not exhibit an endurance limit in fatigue. Moreover, the low hardness of aluminum leads to generally poor wear resistance.

Further important properties, which contribute to the still growing application of aluminum and aluminum alloys, are their high electrical and heat conductivity as well as their good formability. Furthermore, the high chemical affinity of aluminum to oxygen leads to a very acceptable corrosion and oxidation resistance but causes at the same time the high production costs mentioned. To manufacture primary aluminum based on aluminum oxide a costly reduction (mostly smelting electrolytic reduction) is necessary. Bauxite with hydrated forms of aluminum oxide serves as the primary mineralogical source of alumina since its reduction is economically most efficient amongst the different types of aluminum ores.

Al alloys can be very well machined by chip removal, and additions of Pb can prevent the formation of long chips in the case of pure aluminum or soft Al alloys. Joining of Al-based components can be done by all common procedures. Fusion welding is predominantly done by inert gas welding. Adhesion joints are also gaining importance. Aluminum and its alloys do not show a sharp ductile-to-brittle transition temperature; rather they remain ductile even at very low temperatures. Comprehensive treatments of Al-based materials are given in [3.135–137].

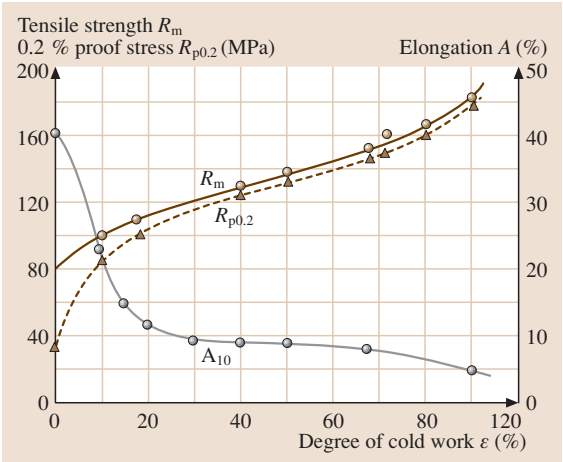


Fig. 3.143 Hardening of Al 99.5 strip (0.15 wt. % Si, 0.28 wt. % Fe) after recrystallization annealing and subsequent cold-rolling (after [3.1])

Table 3.23 The various degrees of purity of pure aluminum [3.135]

Aluminum (%)	Examples (ISO)	Examples (AA)	Designation
99.5000 to 99.7900	A 199.5–A 199.8	1050–1080, 1145	Commercial purity
99.80000 to 99.9490	A 199.8–A 199.95R	1080–1090, 1185	High purity
99.9500 to 99.9959	A 199.95R–A 199.99R	1098, 1199	Super purity
99.9960 to 99.9990	A 199.99R	–	Extreme purity
> 99.9990	–	–	Ultra purity

Table 3.24 Constitution of aluminum alloys

Wrought alloys	
1xxx Commercial pure Al (> 99% Al)	Not aged
2xxx Al–Cu	Age hardenable
3xxx Al–Mn	Not aged
4xxx Al–Si and Al–Mg–Si	Age hardenable if Mg is present
5xxx Al–Mg	Not aged
6xxx Al–Mg–Zn	Age hardenable
7xxx Al–Mg–Zn	Age hardenable
8xxx Other elements (for example Al–Li)	Depends on additions
Casting alloys	
1xx.x Commercial pure Al	Not aged
2xx.x Al–Cu	Age hardenable
3xx.x Al–Si–Cu or Al–Mg–Si	Some are age hardenable
4xx.x Al–Si	Not aged
5xx.x Al–Mg	Not aged
7xx.x Al–Mg–Zn	Age hardenable
8xx.x Al–Sn	Age hardenable
9xx.x (Other elements)	Depends on additions

Pure Aluminum

Commercial-purity aluminum, mainly manufactured by modified Hall–Héroult electrolysis, usually reaches a purity of 99.5–99.8%. On further electrolytic refinement (the three-layer method [3.135]) of commercially pure aluminum or secondary aluminum, superpurity aluminum (99.95–99.99%) can be prepared. Finally, for special purposes, aluminum can be further purified by zone melting to result in extreme purity aluminum of up to 99.99995%. Classification of pure aluminum is given in Table 3.23 of [3.135].

In the annealed condition aluminum possesses only low strength at room temperature. By cold deformation, however, it is possible to improve its strength significantly, whereas the elongation is reduced considerably (Fig. 3.143).

Traditionally, pure aluminum is used in wrought condition for electrical conductors (EC-aluminum). Further important applications of aluminum are as foils for the food processing industries and in packaging practice (alloy 1145), as case components, boxes in tool-building, in the building industry as well as claddings, and to improve resistance to corrosion with heat-treatable Al alloys.

Aluminum Alloys

The major alloying elements of aluminum are copper, manganese, magnesium, silicon, and zinc. Depending on the production route to its final form, aluminum alloys may in principle be divided into wrought alloys and cast alloys. The wrought alloys can be classified into two main groups:

1. Age-hardenable alloys
2. Non-age-hardenable alloys

The nomenclature used for wrought alloys consists of four digits 2xxx-8xxx where the last two digits are the alloy identifier (Table 3.24). The second digit indicates certain alloy modifications (0 stands for the original alloy). A second designation is usually used, and describes the final temper treatment (Table 3.25).

Aluminum responds readily to strengthening mechanisms (Sect. 3.1) such as age hardening, solution hardening, and strain hardening, resulting in 2–30 times higher strength compared with pure aluminum (Table 3.26). Age hardening is the most effective hardening mechanism. It is based on the fact that the solubility of certain elements increases on increasing temperature. In the case of Cu as the alloying element, maximum solubility is reached at about 550 °C (Fig. 3.144). For age hardening the material is solution annealed in the single-phase region, quenched to room or low temperature, and finally age hardened at higher temperatures (100–200 °C) to facilitate the formation of small precipitates. On further age hardening the precipitates continue to grow, resulting in overaging (Fig. 3.144), which is accompanied by a loss in material strength.

Table 3.25 Heat treatments of aluminum alloys

F	As-fabricated (hot worked, forged, cast, etc.)
O	Annealed (in the softest possible condition)
H	Cold worked
	H1x – cold worked only ("x" referes to the amount of cold work and strengthening)
	H-12 – cold work that gives a tensile strength midway between the O and H14 tempers
	H-14 – cold work that gives a tensile strength midway between the O and H18 tempers
	H-16 – cold work that gives a tensile strength midway between the H14 and H18 tempers
	H-18 – cold work that gives about 75% reduction
	H-19 – cold work that gives a tensile strength greater than 2000 psi of that obtained by the H18 temper
	H2x – cold worked and partly annealed
	H3x – cold worked and stabilized at a low temperature to prevent age hardening of the structure
W	Solution treated
T	Age hardened
	T1 – cooled from the fabrication temperature and naturally aged
	T2 – cooled from the fabrication temperature, cold worked, and naturally aged
	T3 – solution treated, cold worked, and naturally aged
	T4 – solution treated and naturally aged
	T5 – cooled from the fabrication temperature and artifically aged
	T6 – solution treated and artifically aged
	T7 – solution treated and stabilized by overaging
	T8 – solution treated, cold worked, and artifically aged
	T9 – solution treated, artifically aged, and cold worked
	T10 – cooled from the fabrication temperature, cold worked, and artifically aged

The strength increase $\Delta\sigma$ is inversely proportional to the separation distance l of the precipitates and is giving in the peak aged condition (Fig. 3.145) by $\Delta\sigma \sim 2Gb/l$ (G – shear modulus; b – Burger vector).

However, on further annealing the precipitates can grow by Ostwald ripening, i. e., small precipitates are consumed and larger particles grow continuously at their expense. This process results in severe strength decrease when the material is exposed to high temperatures during service (Fig. 3.146).

Depending on the alloying additions, different strengthening mechanisms are activated:

- 2xxx: Precipitation of Cu-rich phases allows the formation of high-strength alloys at the expense of weldability. Precipitation from the α-solid so-

Table 3.26 Effect of strengthening mechanisms on the mechanical properties of aluminum alloys (after data in [3.137])

Material	Tensile strength (MPa)	Yield strength (MPa)	(%) Elongation	Yield strength (alloy) / Yield strength (pure)
Pure annealed Al (99.999% Al)	45	17	60	
Commercially pure Al (annealed, 99% Al)	90	34	45	2.0
Solid solution strengthened (1.2% Mn)	110	41	35	2.4
75% cold worked pure Al	165	152	15	8.8
Dispersion strengthened (5% Mg)	290	152	35	8.8
Age hardened (5.6% Zn–2.5% Mg)	570	503	11	29.2

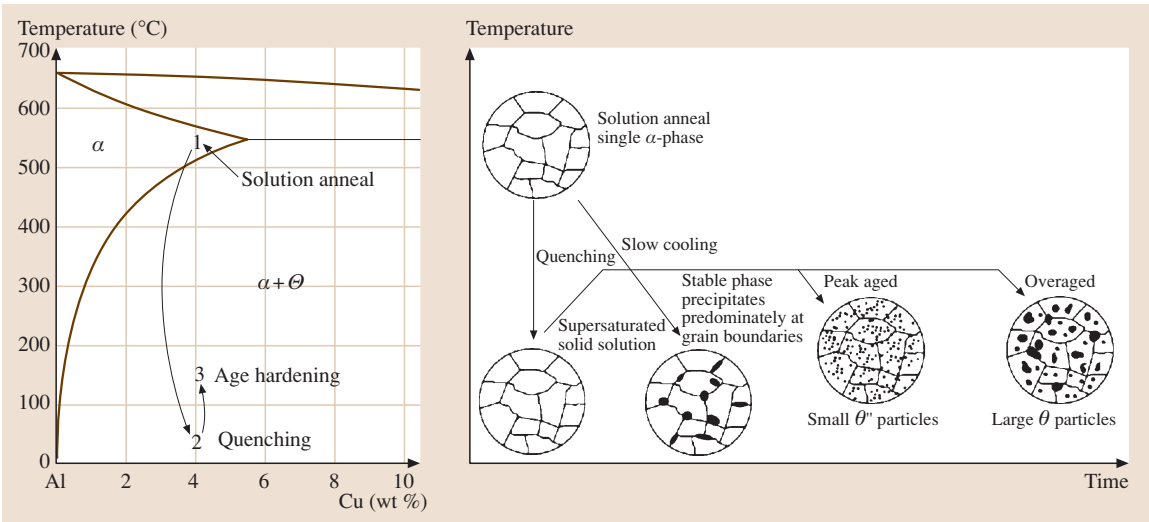


Fig. 3.144 Principle of age hardening in Al-based alloys

lution appears in the following order: $\alpha_{ss} \rightarrow \alpha + GP-I \rightarrow \alpha + GP-II \rightarrow \alpha + \theta' \rightarrow \alpha + \theta$. As shown in Fig. 3.147 the highest strength is reached when coherent GP–II zones have been formed, where Orowan by-passing and precipitation cutting require almost the same amount of energy.

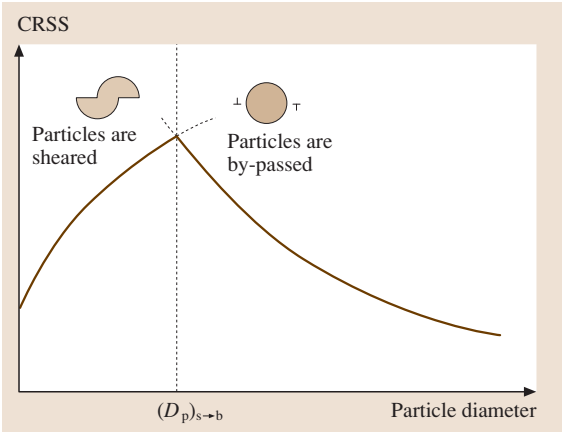


Fig. 3.145 Change in critical resolved shear stress (CRSS - at which dislocations glide freely in single crystals) as a function of particle (precipitate) size. Maximum strength is obtained where dislocation interaction changes from shearing to Orowan by-passing. In the peak aged condition the average particle size corresponds to D_p and particle size variation should be marginal (after [3.135])

- 3xxx: These alloys are single-phase alloys except for the presence of inclusions or intermetallic compounds. Their strength is achieved by solid solution of Mn and is not as high as for 2xxx alloys but they exhibit excellent corrosion and oxidation resistance and are weldable.
- 4xxx: The high strength of this group of wrought alloys is achieved via solid solution of Mg and in the case of the presence of Si by the formation of finely dispersed Mg_2Si particles as well.
- 5xxx: These alloys are strengthened by a solid solution of Mg and contain a second phase: Mg_2Al_3 particles, a hard and brittle intermetallic compound. The corrosion resistance of these alloys is almost comparable to that of pure Al.
- 6xxx: Alloys belonging to this group show a good balance of their properties. They are moderately heat treatable and show moderate corrosion resistance and weldability. $\alpha_{ss} \rightarrow \alpha + GP$ zones $\rightarrow \alpha + \beta\ (Mg_2Si) \rightarrow \alpha + \beta\ (Mg_2Si)$.
- 7xxx: The highest strength levels of commercial aluminum alloys are attained by the members of this group, which are strengthened by addition of Mg and Zn. The alloys are age hardenable by the formation of $MgZn_2$ particles in the following sequence: $\alpha_{ss} \rightarrow \alpha + GP$ zones $\rightarrow \alpha + \eta\ (MgZn_2) \rightarrow \alpha + \eta\ (MgZn_2)$. Most alloys of this group are not weldable. An exemption is the alloy 7005, which is used as a standard alloy for bike frames.

Table 3.27 Mechanical properties and applications of selected aluminum alloys (after [3.137, 138])

Alloy	Chemical composition	Condition	Tensile strength (MPa)	Yield strength (MPa)	Elongation (%)	Typical applications
3003	1.2 Mn	Annealed (-O) Half-hard (-H14)	110 150	40 145	30 8	Pressure vessels, builders' hardware, sheet metal work
5052	2.5 Mg, 0.25Cr	Annealed (-O) Half-hard (-H32)	195 230	90 195	25 12	Sheet metal work, hydraulic tubes, appliances
2024	4.4 Cu, 1.5 Mg, 0.6 Mn	Annealed (-O) Heat-treated (-T6)	220 442	97 345	12 5	Truck wheels, screw machine product, aircraft structures
6061	1.0 Mg, 0.6 Si, 0.27 Cu, 0.2 Cr	Annealed (-O) Heat-treated (-T6)	125 310	55 275	25 12	Heavy-duty structures requiring good corrosion resistance, pipelines
7075	5.6 Zn, 2.5 Mg, 1.6 Cu, 0.23 Cr	Annealed (-O) Heat-treated (-T6)	230 570	105 505	17 11	Aircraft and other structures

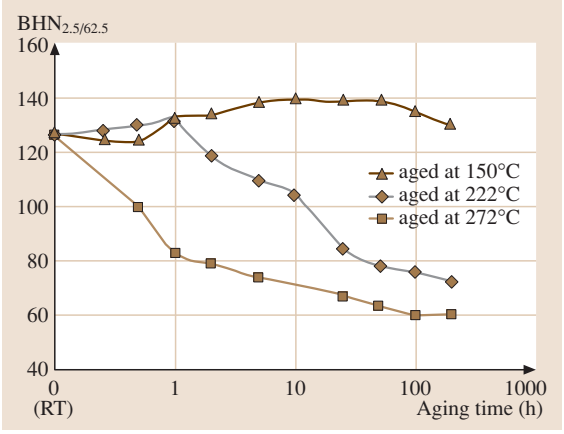


Fig. 3.146 Room-temperature hardness (BHN 2.5/62.5) of T6 age-hardened Al_6Si_4Cu (A319) after static annealing at different temperatures for up to 200 h (after [3.139])

Mechanical properties and some typical applications of selected aluminum alloys are given in Table 3.27.

Cast aluminum alloys are developed to have good fluidity and feeding ability during casting. Their designation (Table 3.24) is based on three major digets 2xx-9xx, giving the alloy group, and a further digit following a dot, which indicates the material form (casting/ingot). The most widely used casting alloys belong to the 300 series (319 and 356), where hardening is done by Cu or Mg_2Si precipitation. Examples of casting alloys are giving in Table 3.28.

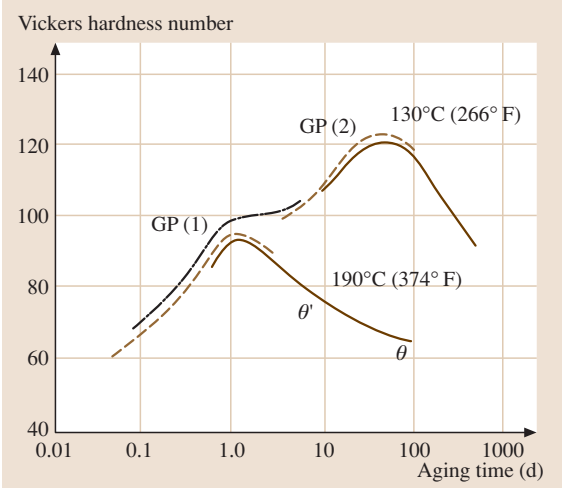


Fig. 3.147 Room temperature hardness (Vickers) of 2000 series alloy after static annealing at different temperatures (after [3.138])

Table 3.28 Selected cast aluminum alloys and their mechanical properties (after data in [3.140], see also [3.137])

Alloy	Chemical composition	Tensile strength (MPa)	Yield strength (MPa)	Elongation (%)	Casting process
201-T6	4.5% Cu	483	434	7	Sand
319-F	6% Si 3.5% Cu	186	124	2	Sand
		234	131	2.5	Permanent mold
356-T6	7% Si 0.3% Mg	228	165	3.5	Sand
		262	186	5	Permanent mold
380-F	8.5% Si 3.5% Cu	317	156	3.5	Permanent mold
384-F	11.2% Si 4.5% Cu 0.6% Mg	331	165	2.5	Permanent mold
390-F	17% Si 4.5% Cu 0.6% Mg	283	241	1	Die casting
443-F	5.2% Si	131	55	8	Sand
		159	62	10	Permanent mold
		228	110	9	Die casting
413-F	12% Si	296	145	2.5	Die casting
518-F	8% Mg	310	193	7	Sand
713-T5	7.5% Zn 0.7% Cu 0.35% Mg	207	152	4	Sand
850-T5	6.2% Sn 1% Ni 1% Cu	159	76	10	Sand

3.7.3 Magnesium and Its Alloys

General Properties

Magnesium is the lightest structural metal with a density close to that of polymers (plastics). It is therefore not surprising that Mg alloys are especially found in applications where the weight of a workpiece is of paramount importance, as generally is the case in the transportation industry. In recent years magnesium cast alloys have particularly becoming increasingly important and have partly replaced well-established Al-based alloys. The main reason is the excellent die-filling characteristics of magnesium, which allows large, thin-walled, and unusually complex castings to be produced economically. Magnesium can be cast with thinner walls (1–1.5 mm) than plastics (2–3 mm) or aluminum (2–2.5 mm) and, by designing appropriately located ribs, the stiffness disadvantage of magnesium versus aluminum can be compensated without increasing the wall thickness of an overall magnesium part.

Further positive properties to be noted are the excellent machinability, high thermal conductivity (Sect. 3.4), and the good weldability. However, Mg alloys suffer from poor corrosion resistance and the manufacturing costs are comparatively high. With its hexagonal close-packed crystal structure the room-temperature deformation behavior of Mg alloys is moderate, resulting in poor cold workability. Thus, all current applications are manufactured through casting. Furthermore, Mg is a very reactive metal and readily oxidizes when exposed to air. Since pure Mg is only of minor importance for structural applications it appears almost always in the alloyed condition with additions such as Al and Zn. A comprehensive treatment of Mg and its alloys is given in [3.141].

Magnesium Alloys

Major alloying elements of Mg are Al, Zn, and Mn, while elements such as Sn, Zr, Ce, Th, and B are occasionally of importance. Impurities in Mg alloys are commonly Cu, Fe, and Ni. Mg designation is based on the main alloying elements (such as AZ for aluminum and zinc) followed by the amount of additives and a letter that indicates the amount of variations with

Table 3.29 Designation of Mg alloys

1.	Two letters which indicate the major alloying additions A–Al; Z–Zn; M–Mn; K–Zr; T–Sn; Q–Ag; C–Cu; W–Y; E–rare earths
2.	Two or three numbers which indicate the nominal amounts of alloying elements (rounded off to the nearest percent)
3.	A letter which describes variation to the normal alloy
4.	If needed, the temper treatment according to Table 3.30

Table 3.30 Temper designations (after [3.1])

General designations			
F	As fabricated		
O	Annealed. recrystallized (wrought products only).		
H	Strain-hardened		
T	Thermally threated to produce stable tempers other than F, O, or H.		
W	Solution heat-treated (unstable temper).		
Subdivisions of H		**Subdivisions of T**	
H1, Plus one or more digits	Strain only	T2	Annealed (cast products only)
H2, Plus one or more digits	Strain-hardened and then partially annealed	T3	Solution heat-treated and cold worked
H3, Plus one or more digits	Strein-hardened and then stabilized	T4	Solution heat-treated
		T5	Artificial aged only
		T6	Solution heat-treated and artificial aged
		T7	Solution heat-treated and stabilized
		T8	Solution heat-treated, cold worked, and artificial aged
		T9	Solution heat-treated, artificial aged, and cold worked
		T10	Artificial aged and cold worked

Table 3.31 General effects of alloying elements in magnesium materials (after [3.1], see also [3.141–143])

Series	**Alloying elements**	**Melting and casting behavior**	**Mechanical and technological properties**
AZ	Al, Zn	Improve castability; tendency to microproporosity; increase fluidity of the melt; refine weak grain	Solid-solution hardener; precipitation hardening at low temperatures (< 120 °C); improve strength at ambient temperatures; tendency to brittleness and hot shortness unless Zr is refinded
QE	Ag, rare earths	Improve castability; reduce microporosity	Solid-solution and precipitation hardening at ambient and elevated temperatures; improve elevated-temperature tensile and creep properties in the presence of rare-earth metals
AM	Al, Mn	Improve castability; tendency to microporosity; control of Fe content by precipitating Fe – Mn compound; refinement of precipitates	Solid-solution hardener; precipitation hardening at low temperatures (< 120 °C); increase creep resistivity
AE	Al, rare earth	Improve castability; reduce microporosity	Solid-solution and precipitation hardening at ambient and elevated temperatures; improve elevated-temperature tensile and creep properties; increase creep resistivity
AS	Al, Si	Tendency to microporosity; decreased castability; formation of stable silicide alloying elements; compatible with Al, Zn, and Ag; refine week grain	Solid-solution hardener, precipitation hardening at low temperatures (< 120 °C); improves creep properties
WE	Y, rare earths	Grain refining effect; reduce microporosity	Improve elevated-temperature tensile and creep properties; solid-solution and precipitation hardening at ambient and elevated temperatures

respect to the normal alloy (Table 3.29). When referring to mechanical properties it is useful to indicate the temper treatment as well (Table 3.30). The alloy AZ91A, for example, is a Mg-based alloy with nominally about 9% Al and 1% Zn, while the letter A indicates that only minor changes to the normal alloy were carried out.

Table 3.32 Typical tensile properties and characteristics of selected cast Mg alloys (after [3.1], see also [3.141–143])

ASTM designation	Condition	Tensile properties 0.2% proof stress (MPa)	Tensile strength (MPa)	Elogation to fracture (%)	Characteristics
AZ63	As-sand cast	75	180	4	Good room-temperature strength and ductility
	T6	110	230	3	
AZ81	As-sand cast	80	140	3	Tough, leaktight casting with 0.0015 Be, used for pressure die casting
	T4	80	220	5	
AZ91	As-sand cast	95	135	2	General-purpose alloy used for sand and die casting
	T4	80	230	4	
	T6	120	200	3	
	As-chill cast	100	170	2	
	T4	80	215	5	
	T6	120	215	2	
AM50	As-die cast	125	200	7	High-pressure die casting
AM20	As-die cast	105	135	10	Good ductility and impact strength
AS41	As-die cast	135	225	4.5	Good creep properties up to 150 °C
AS21	As-die cast	110	170	4	Good creep properties up to 150 °C
ZK51	T5	140	253	5	Sand casting, good room-temperature strength and ductility
ZK61	T5	175	275	5	As for ZK51
ZE41	T5	135	180	2	Sand casting, good room-temperature strength, improved castability
ZC63	T6	145	240	5	Pressure-tight casting, good elevated-temperature strength, weldable
EZ33	Sand cast T5	95	140	3	Good castability, pressuretight, weldable, creep resistant up to 250 °C
	Chill cast T5	100	155	3	
HK31	Sand cast T6	90	185	4	Sand casting, good castability, weldable, creep resistant up to 350 °C
HZ32	Sand or chill cast T5	90	185	4	As for HK31
QE22	Sand or chill cast T6	185	240	2	Pressuretight and weldable, high proof stress up to 250 °C
QH21	As-sand cast T6	185	240	2	Pressuretight, weldable, good creep resistance and stressproof to 300 °C
WE54	T6	200	285	4	High strength at room and elevated temperatures, good corrosion resistance
WE43	T6	190	250	7	Weldable

An overview of the general effect of certain alloying additions is given in Table 3.31 [3.1, 141–143]. The addition of up to 10% aluminum (Mg−Al alloys) increases the strength (age hardenable), castability, and corrosion resistance. During precipitation heat treatment the intermetallic phase $Mg_{17}Al_{12}$ is formed, and is usually not finely distributed enough to lead to a strong strengthening effect. The supplementary addition of zinc (Mg−Al−Zn alloys) improves the strength of Mg−Al alloys by refining the precipitates and by solid-solution strengthening. The frequently used alloy AZ91, for example, offers yield strength and ductility levels which are comparable to its aluminum counterpart A380. However, in terms of high-temperature creep resistance (application limited to about 125 °C), fatigue strength, and corrosion resistance the alloy AZ91 is inferior to Al alloys. Its application is therefore restricted to nonstructural components such as

brackets, covers, cases, and housings. With the development of the low-impurity AZ91D alloy (Fe < 0.005%; Ni < 0.001%, Cu < 0.015%) corrosion resistance has been improved dramatically. For structural applications where crashworthiness is important such as instrument panels, steering systems, and seating structures, magnesium die-cast alloys with small amounts of manganese (Mg−Al−Mn alloys) such as AM50 or AM60 are used as they offer higher ductility (elongation to failure: 10–15%). These alloys are less expansive than Al alloys. The poor high-temperature creep resistance of Mg alloys is the reason why these alloys are rarely found in automotive powertrains. The high operating temperatures (transmission cases: 175 °C, engine blocks: > 200 °C) are already a challenge for aluminum alloys, which have higher creep resistance. For this application Mg alloys containing rare-earth elements are under development to improve creep resistance by precipitation strengthening (some examples such as QE22 and WE43 are given in Tables 3.32,3.33 [3.1, 141–143]).

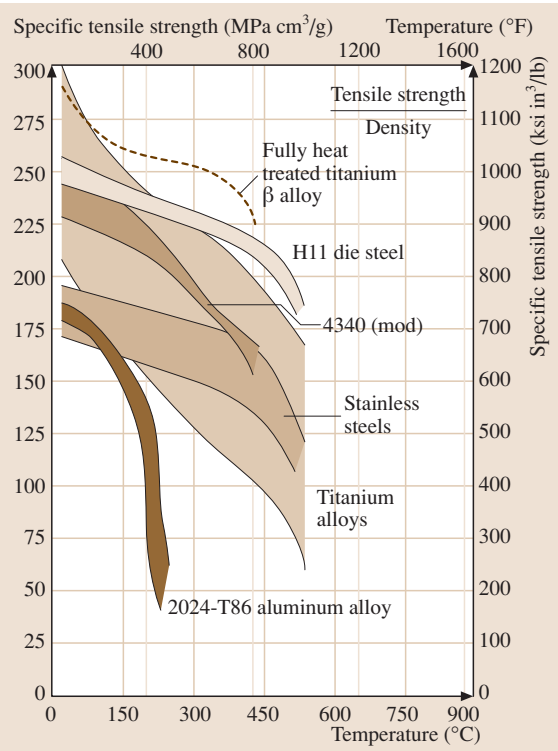


Fig. 3.148 High-temperature properties of Ti alloys compared with those of steels and aluminum alloys (after [3.144])

3.7.4 Titanium and Its Alloys

General Properties

With a density below 5 g/cm^3 (Table 3.13) titanium, like aluminum and magnesium, belongs to the light structural materials. However, in contrast to Al and Mg it also offers a high melting point of $T_s = 1670\,°C$, which is even higher than that of pure iron. Furthermore, amongst the materials which are under consideration for light structural constructions, Ti-based alloys have the highest specific strength (Table 3.13) and excellent corrosion resistance in oxidizing acids, chloride media, and most neutral environments. Titanium is well known for its biocompatibility, low thermal conductivity ($\kappa = 21\,W\,m^{-1}\,K^{-1}$), and low thermal expansion coefficient ($\lambda = 8.9\times10^{-6}\,K^{-1}$). Beside a slightly lower specific stiffness compared with iron-based materials the most annoying disadvantage of Ti and its alloys is the high manufacturing cost, which is about six times that of aluminum and ten times that of stainless steels. Therefore, Ti alloys are primary used in areas where strength-to-weight ratio and elevated-temperature properties are of prime importance, i. e., in aerospace applications (compare Fig. 3.148 [3.144]).

As an allotropic material, Ti undergoes a structural phase transformation at 883 °C from a (almost) closed-packed hexagonal structure (α) to a body-centered cubic high-temperature phase (β). By alloying additions and applying appropriate heat treatments titanium alloys can be age hardened to form a two-phase ($\alpha+\beta$) alloy. The solubility of interstitials such as O, N, C, and H is very high, allowing on the one hand strength to be increased by solid-solution hardening, but also leading on the other hand to a severe reduction of toughness when the material is penetrated by gases. Comprehensive treatments on Ti and its alloys can be found in [3.142, 145, 146].

Commercially Pure and Low-Alloy Grades of Titanium

Commercially pure grades of titanium in the purity range 99.0–99.5% can actually be considered as α-phase alloys since they contain certain levels of O, N, C, and Fe, resulting from the manufacturing process. Furthermore, oxygen may be added deliberately for solid-solution strengthening. Interstitials in titanium are very effective strengtheners; a 0.1% oxygen equivalent (% O equivalent = % O + 2% N + 0.67% C) in pure titanium increases strength by about 120 MPa. However, interstitials counteract fracture toughness; some applications, especially at low temperatures, may there-

Table 3.33 Typical tensile properties and characteristics of selected wrought Mg alloys (after [3.1], see also [3.141–143])

ASTM designation	Condition	Tensile properties 0.2% proof stress (MPa)	Tensile strength (MPa)	Eloagation to fracture (%)	Characteristics
M1	Sheet, plate F	70	200	4	Low- to medium-strength alloy, weldable, corrosion resistant
	Extrusion F	130	230	4	
	Forgings F	105	200	4	
AZ31	Sheet, plate O	120	240	11	Medium-strength alloy, weldable, good formabilility
	H24	160	250	6	
	Extrusion F	130	230	4	
	Forging F	105	200	4	
AZ61	Extrusion F	105	260	7	High-strength alloy, weldable
	Forging F	160	275	7	
AZ80	Forging T6	200	290	6	High-strength alloy
ZM21	Sheet, plate O	120	240	11	Medium-strength alloy, good formability, good damping capacity
	H24	165	250	6	
	Extrusions	155	235	8	
	Forgings	125	200	9	
ZMC711	Extrusions T6	300	325	3	High-strength alloy
LA141	Sheet, plate T7	95	115	10	Ultra-lightweight (S.G. 1.35)
ZK61	Extrusion F	210	185	6	High-strength alloy
	T5	240	305	4	
	Forging T5	160	275	7	
HK31	Sheet, plate H24	170	230	4	High-creep sesistance to 350 °C, weldable
	Extrusion T5	180	255	4	
HM21	Sheet, plate T8	135	215	6	High-creep sesistance to 350 °C, weldable after short-time exposure to 425 °C
	T81	180	255	4	
	Forging T5	175	225	3	
HZ11	Extrusion F	120	215	7	Creep resistance to 350 °C, weldable

Table 3.34 Chemical composition and the mechanical properties of commercial pure and low-alloy grades of titanium (from [3.1])

O (wt.%)	Tensile strength R_m (MPa)	Yield strength $R_{p0.2}$	Fracture strain A_{10} (%)	Standard grade [a] cp	Standard grade [a] low alloyed
0.12	290–410	> 180	> 30	Grade 1	Pd: grade 11
0.18	390–540	> 250	> 22	Grade 2	Pd: grade 7 Ru: grade 27
0.25	460–590	> 320	> 18	Grade 3	Ru: grade 26
0.35	540–740	> 390	> 16	Grade 4	
0.25	> 480	> 345	> 18		Ni + Mo: grade 12

[a] ASTM B265, ed 2001; N_{max}: 0.03 wt. %; C_{max}: 0.08 wt. %; H_{max}: 0.015 wt. %

fore require titanium grades with extra-low interstitials (ELI). While having an hcp structure Ti exhibits surprisingly high room-temperature ductility and can be cold-rolled to > 90% without crack formation. This behavior is attributed to the relative ease of activating slip systems and the availability of twinning planes in the crystal lattice. The chemical composition and the mechanical properties of commercial pure and low-alloy grades of titanium are given in Table 3.34.

Titanium Alloys

Alloying additions, which are usually added to improve the mechanical properties of Ti influence the phase stability in a different manner. The low-temperature

Table 3.35 Alloying elements in Ti alloys [3.142, 144, 145]

Alloying element	Range (approx.) (wt.%)	Effect on structure
Carbon, oxygen, nitrogen	–	α stabilizer
Aluminum	2–7	α stabilizer
Tin	2–6	α stabilizer
Vanadium	2–20	β stabilizer
Molybdenum	2–20	β stabilizer
Chromium	2–12	β stabilizer
Copper	2–6	β stabilizer
Zirconium	2–8	α and β strengtheners
Silicon	0.05–1	Improves creep resistance

Table 3.36 Chemical composition and mechanical properties of Ti-based alloys at room temperature (minimum values) (after [3.1])

Alloy composition [a]	Alloy types	Tensile strength R_m (MPa)	Yield strength $R_{p0.2}$ (MPa)	Density ϱ (g/cm^3)	Young's modulus E (GPa)	Main property	Standard grade [b]
Ti5Al2.5Sn	α	830	780	4.48	110	High strength	
Ti6Al2Sn4Zr2MoSi	near α	900	830	4.54	114	High-temperature strength	3.7145
Ti6Al5Zr0.5MoSi	near α	950	880	4.45	125	High-temperature strength	3.7155
Ti5.8Al4Sn3.5Zr0.7Nb 0.5Mo0.2Si0.05C	near α	1030	910	4.55	120	High-temperature strength	
Ti6Al4V	$\alpha+\beta$	900	830	4.43	114	High strength	
Ti4Al4Mo2Sn	$\alpha+\beta$	1100	960	4.60	114	High strength	3.7185
Ti6Al6V2Sn	$\alpha+\beta$	1030	970	4.54	116	High strength	3.7185
Ti10V2Fe3Al	near β	1250	1100	4.65	103	High strength	
Ti5V3Cr3Sn3Al	β	1000	965	4.76	103	High strength; cold formability	
Ti3Al8V6Cr4Zr4Mo	β	1170	1100	4.82	103	High corrosion; resistance	
Ti15Mo3Nb3AlSi	β	1030	965	4.94	96	High corrosion; resistance	

[a] Figure before chemical symbol denotes nominal wt.%
[b] According to DIN 17851, ASTM B 265 ed. 2001

hexagonal α-phase is stabilized by the impurities O, N, and C as well as by Al and Sn (Table 3.35), whereas elements such as V, Mo, and Cr expand the β-phase stability region (the Ti-rich part of the Ti−Al and the Ti−Mo phase diagram are shown in Fig. 3.149 [3.147]).

By varying the alloying content pure α- or β-phase alloys can be stabilized at room temperature as well as a mixture of both phases. The α-phase Ti alloys have a high solid solubility at room temperature and are weldable. The most widely used α-Ti alloy is Ti-5Al-2.5Sn (Table 3.36). While offering the highest strength levels of the Ti alloys and the ability of cold working, the usage of β-phase alloys is rather limited compared with pure α- or $\alpha+\beta$-alloys. Besides costs, the reasons for this include the higher density, caused by the addition of V or Mo, the low ductility in the high-strength condition, and the poor fatigue performance in thick sections, which is caused by segregations at grain boundaries. The most widely used group (about 60%) of Ti alloys are two-phase $\alpha+\beta$-alloys, with Ti-6Al-4V being the most prominent representative. These alloys are heat treatable and allow large variations of the mi-

Table 3.37 Typical applications of two-phase $\alpha + \beta$-alloys (after [3.144])

Alloy composition	Condition	Typical applications
6% Al 4% V	Annealed; solution + age	Rocket motor cases; blades and disks for aircraft turbines and compressors; structural forgings and fasteners; pressure vessels; gas and chemical pumps; cryogenic parts; ordenance equipment; marine components; steam-turbine blades.
6% Al, 4% V, low O_2 Sn	Annealed	High-pressure cryogenic vessels operating down to −196 °C.
6% Al, 6% V, 2% Sn	Annealed; solution + age	Rocket motor cases; ordenance components; structural aircraft parts and landing gears; responds well to heat treatments; good hardenability.
7% Al, 4% Mo	Solution + age	Airframes and jet engine parts for operation at up to 427 °C; missile forgings; ordenance equipment.
6% Al, 2% Sn, 4% Zr, 6% Mo	Solution + age	Components for advanced jet engines.
6% Al, 2% Sn, 2% Zr, 2% Mo, 2% Cr, 0.25% Si	Solution + age	Strength, fracture toughness in heavy sections; landing-gear wheels.
10% V, 2% Fe, 3% Al	Solution + age	Heavy airframe structural components requiring toughness at high strengths.
8% Mo	Annealed	Aircraft sheet components, structural sections and skins; good formability, moderate strength.
3% Al, 2.5% V	Annealed	Aircraft hydraulic tubing, foil, combines strength, weldability, and formability.

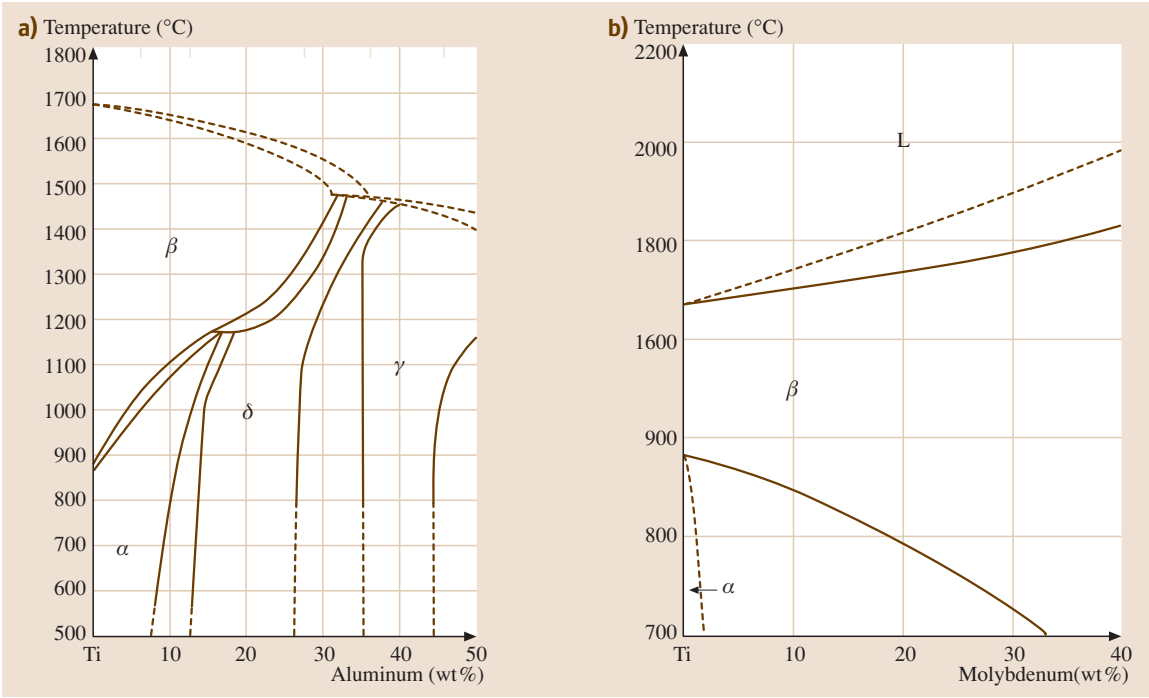


Fig. 3.149a,b Ti-rich part of the (a) Ti−Al and (b) Ti−Mo phase diagrams, showing the effect of Al as α-phase and Mo as β-phase stabilizers, respectively (after [3.147])

crostructure by altering the cooling and heat-treatment conditions. Some typical applications of $\alpha + \beta$-alloys and the conditions of their usage are given in Table 3.37.

Two-Phase Intermetallic Ti–Al Alloys

With a density as low as 3.5 g/cm^3 and a specific stiffness of as high as 45 GPa cm^3 g (steel: 27 GPa cm^3 g)

Table 3.38 Typical applications and properties of Ni-based alloys and superalloys

Material	Tensile strength (MN/m^2)	Yield strength (MN/m^2)	Elongation (%)	Strengthening mechanism	Application
Pure Ni (99.9% Ni)	345	110	45	Annealed	Corrosion resistance
	655	620	4	Cold-worked	Corrosion resistance
Ni-Cu alloys					
Monel 400 (Ni-31.5% Cu)	540	270	37	Annealed	Valves, pumps, heat exchangers
Monel K-500 (Ni-29.5% Cu-2.7% Al-0.6% Ti)	1030	760	30	Aged	Shafts, springs, impellers
Ni superalloys					
Inconel 600 (Ni-15.6% Cr-8% Fe)	620	200	49	Carbides	Heat-treatment equipment
Hastelloy B-2 (Ni-28% Mo)	900	415	61	Carbides	Corrosion resistance
DS-Ni (Ni-2% ThO_2)	490	330	14	Dispersion	Gas turbines
Fe-Ni superalloys					
Incoloy 800 (Ni-46% Fe-21% Cr)	615	258	37	Carbides	Heat exchangers
Co superalloys					
Stellite 6B (60% Co-30% Cr-4.5% W)	1220	710	4	Carbides	Abrasive wear resistance

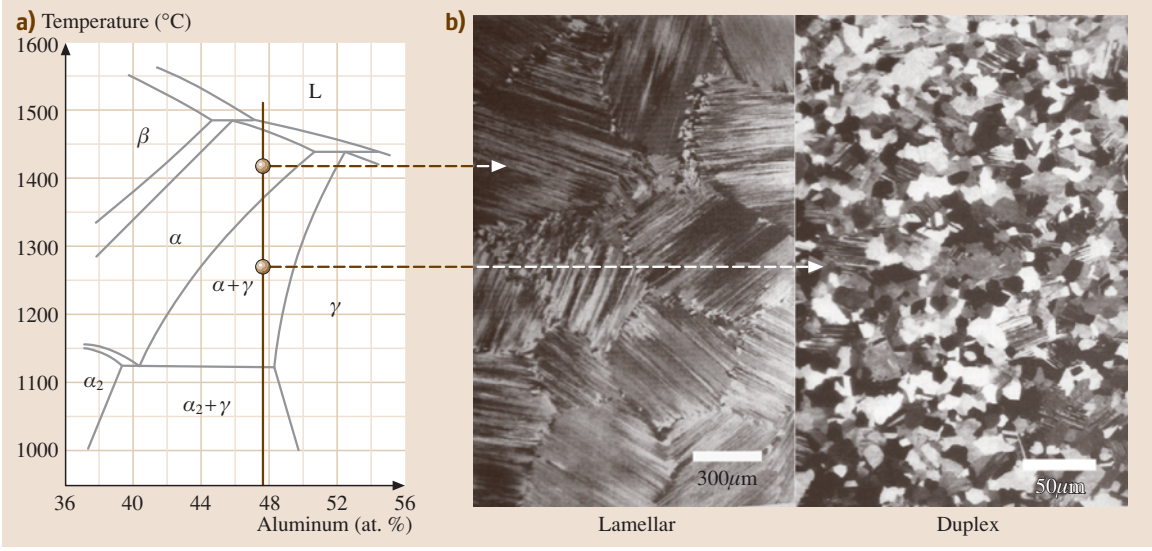


Fig. 3.150 **(a)** Partial Ti−Al phase diagram near the stoichiometric TiAl composition. The marks indicate the heat treatment temperature; **(b)** resulting microstructures after heat treatment and cooling to room temperature (solid-state transitions) (after [3.149])

intermetallic compounds in the Ti−Al system are attractive candidates for high-temperature applications (see, for example, [3.148]), mainly because of their unique combination of very low density and high melting point (above 1350 °C). The most promising alloys are Ti-rich two-phase (γ-TiAl and α_2-Ti_3Al) Ti–Al alloys with lamellar or duplex microstructures, which are adjusted by adequate heat treatments (Fig. 3.150). Additions such as Cr and Nb lead to further enhancement of the mechanical properties and allow elongations of up to about 3%, and room-temperature toughness values of 10–35 MPa$\sqrt{\text{m}}$ underline their principal suitability as constructional materials.

Ti–Ni Shape-Memory Alloys

Ti–Ni alloys are the most prominent representatives of shape-memory alloys [3.150]. The shape-memory effect, allowing the return of a highly deformed material to its starting shape, is based on a reversible martensitic transformation. In the case of Ti–Ni this transition is provided by a transition of the cubic, high-temperature B2 structure to the monoclinic low-temperature B19 structure upon cooling or by deformation. The transformation start (M_s) temperatures can be varied between −200 and 110 °C by altering the Ni content. Therefore, the transformation can be reversed either on heating or on releasing the stress isothermally. For further reading see [3.1, 150].

3.7.5 Ni and Its Alloys

General Properties

Due to its similarity to Fe with respect to the most relevant physical (and chemical) properties such as density, Young's modulus, melting point, thermal conductivity, and CTE (Sect. 3.4), it is straightforward to conclude that nickel is one of the major alloying elements in Fe-based alloys. Since Ni possesses a fcc crystal structure it stabilizes the austenite in Fe-based alloys at higher concentrations. In fact, over 60% of the annual Ni con-

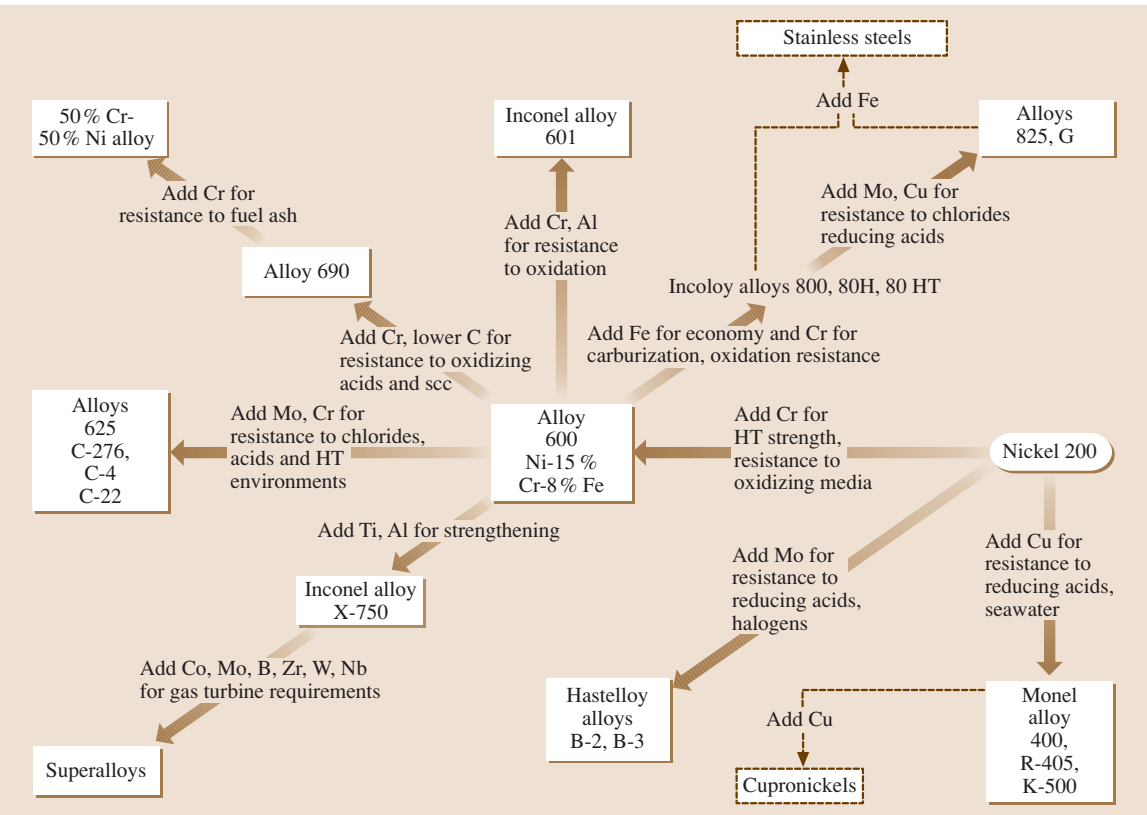


Fig. 3.151 Effects of alloying additions on the corrosion resistance of nickel alloys (HT denotes high temperature) (after [3.1])

sumption is devoted to alloying of stainless steels and a further 10% is used in (ferritic) alloy steels.

Nickel forms extensive solid solutions with many other elements: complete solid solutions with Fe and Cu (such as those exemplified with the phase diagrams in Figs 3.30,3.31) and limited solid solutions with < 35 wt. % Cr, < 20 wt. % Mo, < 10 wt. % Al, Ti, to mention the most important ones. Based on the fcc crystal structure Ni and its solid solutions show high ductility, fracture toughness, and formability.

Alloys of Ni−Fe show ferromagnetism over a wide range of compositions which, in combination with other intrinsic properties, gives rise to alloys with soft magnetic [3.59] and controlled thermal expansion properties (*Invar alloy*, Sect. 3.4.1). Ti–Ni shape-memory alloys are briefly discussed in Sect. 3.7.4. Finally, nickel plating is widely used for decorative applications. Most frequently, electroless deposition of either nickel–phosphorous or nickel–boron binary solutions is carried out by autocatalytic reduction of Ni ions from aqueous solutions. For more details see [3.151].

Besides these functional applications, structural applications of nickel and its alloys can be essentially grouped into two categories, namely:

1. Corrosion-resistant alloys
2. High-temperature alloys

as will be described briefly in the following two subsections. A survey on commonly used alloying additions in nickel and their effects on properties and applications is shown in Fig. 3.151.

Corrosion-Resistant Alloys

The main application of commercially pure nickel is to combine corrosion resistance with outstanding formability. The 200 alloy series typically contains minor amounts of less than 0.5 wt. % Cu, Fe, Mn, C, and Si. According to Fig. 3.148 the intrinsically good corrosion resistance of nickel 200 can be substantially improved by high alloying in solid solution with

- Cu for increased resistance against seawater and reducing acids, leading to the Monel alloys (e.g., 400, K-500)
- Mo for increased resistance against reducing acids and halogens, leading to the Hastelloy alloys (B2, B3)
- Cr for increased high-temperature strength and resistance to oxidizing media, leading to alloy 600 (which also possesses about 8 wt. % Fe, mainly for economical reasons)

Alloy 600 can be considered as the base alloy for a series of further high-alloyed Ni-base alloys for various applications in aggressive environments, as displayed in Fig. 3.151. An extensive compilation of chemical compositions and mechanical properties may be found in [3.1] while some typical examples for Ni alloys are listed in Table 3.38 together with their corresponding field of application.

Ni-Based Superalloys

The term *superalloy* is generally used for metallic alloy systems which may operate under structural loading

Table 3.39 Compositions, microstructures, and properties of representative Co-bonded cemented carbides (after [3.1] p. 279)

Nominal composition	Grain size	Hardness (HRA)	Density ($g\,cm^{-3}$)	($oz\,in^{-3}$)	Transverse strength (MPa)	(ksi)	Compressive strength (MPa)	(ksi)
97WC-3Co	Medium	92.5–93.2	15.3	8.85	1590	230	5860	850
94WC-6Co	Fine	92.5–93.1	15.0	8.67	1790	260	5930	860
	Medium	91.7–92.2	15.0	8.67	2000	290	5450	790
	Coarse	90.5–91.5	15.0	8.67	2210	320	5170	750
90WC-10Co	Fine	90.7–91.3	14.6	8.44	3100	450	5170	750
	Coarse	87.4–88.2	14.5	8.38	2760	400	4000	580
84WC-16Co	Fine	89	13.9	8.04	3380	490	4070	590
	Coarse	86.0–87.5	13.9	8.04	2900	420	3860	560
75WC-25Co	Medium	83–85	13.0	7.52	2550	370	3100	450
71WC-12.5TiC -12TaC-4.5Co	Medium	92.1–92.8	12.0	6.94	1380	200	5790	840
72WC-8TiC -11.5TaC-8.5Co	Medium	90.7–91.5	12.6	7.29	1720	250	5170	750

[a] Based on a value of 100 for the most abrasion-resistant material

Table 3.39 (cont.)

Nominal composition	Modulus of elasticity (GPa)	(10^6 psi)	Relative abrasion resistance [a]	Coefficient of thermal expansion (μm/m K) at 200 °C (390°F)	at 1000 °C (1830°F)	Thermal conductivity (W/m K)
97WC-3Co	641	93	100	4.0	–	121
94WC-6Co	614	89	100	4.3	5.9	–
	648	94	58	4.3	5.4	100
	641	93	25	4.3	5.6	121
90WC-10Co	620	90	22	–	–	–
	552	80	7	5.2	–	1.12
84WC-16Co	524	76	5	–	–	–
	524	76	5	5.8	7.0	88
75WC-25Co	483	70	3	6.3	–	71
71WC-12.5TiC -12TaC-4.5Co	565	82	11	5.2	6.5	35
72WC-8TiC -11.5TaC-8.5Co	558	81	13	5.8	6.8	50

conditions at *elevated temperatures* above 1200°F (or around 650 °C correspondingly). Note, that this is synonymous with operating under creep conditions since $T_{app} > 0.5T_m$ (Sect. 3.3.2). As displayed in Fig. 3.152 superalloys can be grouped into three main subcategories according to the strengthening mechanism (Sect. 3.1.2) employed. One distinguishes:

1. Solid-solution-strengthened iron, nickel, and cobalt alloys
2. Carbide-strengthened cobalt alloys (see next section) and most prominently
3. Precipitation-strengthened nickel and nickel–iron alloys

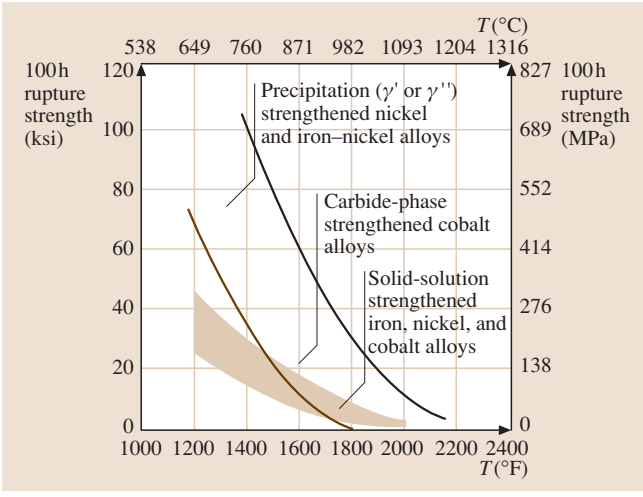


Fig. 3.152 Temperature dependence of the 100 h stress-rupture characteristics of wrought superalloys (after [3.1])

The paramount importance of the latter group regarding high-temperature creep strength stems from alloying with Al and Ti, which leads to the formation of a coherent ordered intermetallic γ'-$Ni_3(Al, Ti)$ phase. This $L1_2$ crystal structure has a superlattice structure with regard to the disordered fcc structure of the γ phase. The binary Al−Ni phase diagram (Fig. 3.153) clearly demonstrates that the γ' phase is stable up to its melting point close to 1400 °C. Since the γ phase exhibits a decreasing solubility for Al with decreasing temperature, precipitation strengthening by age hardening can be carried out in analogy to the Al−Cu system (Fig. 3.143). Maximum fractions of γ' phase exceed 60% by volume in single-crystal cast alloys such as CMSX 4 (Fig. 3.154). The pronounced hardening effect of the precipitate phase is mainly due to the γ/γ' lattice mismatch which causes dislocation reactions at the interphase and an anomalous temperature dependence of strength of the γ' phase [3.17]. The strength anomaly is also the reason why forming and machining operations of wrought superalloys require special attention and tools and restrict γ' volume fractions to below 50%.

Another major alloying element is carbon which forms carbides (of type MC, M_7C_3, $M_{23}C_6$, M_6C) with Ti, Cr, Nb, Mo, and W in order to stabilize the microstructure (grain structure) against creep deformation. The latter *heavy* elements are added also for γ matrix solid-solution strengthening since they segregate preferentially there. This effect has been further accen-

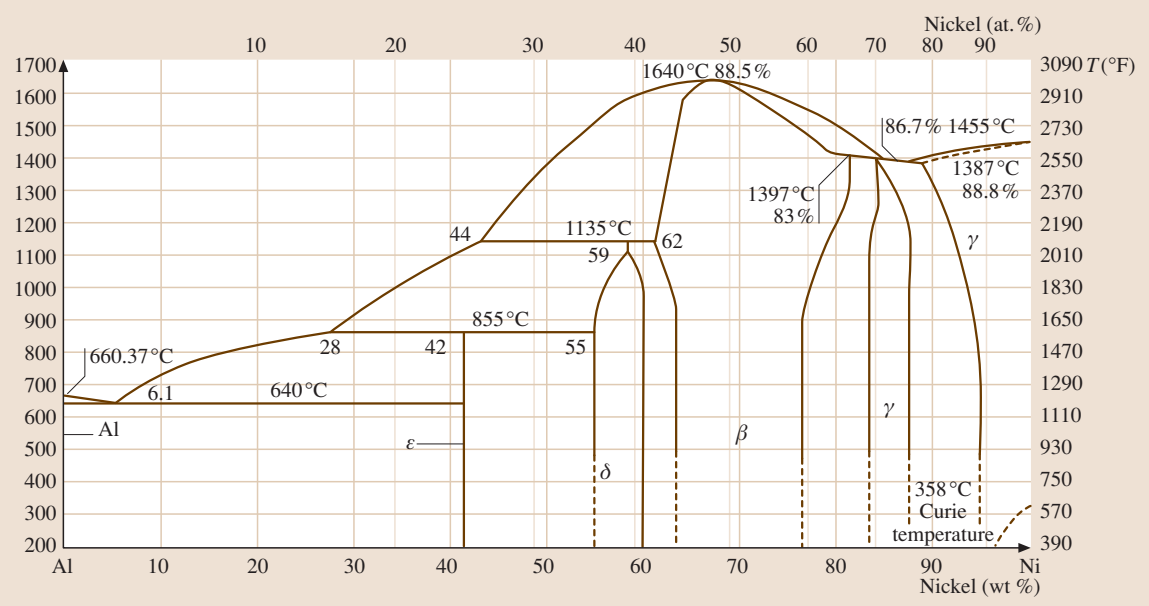
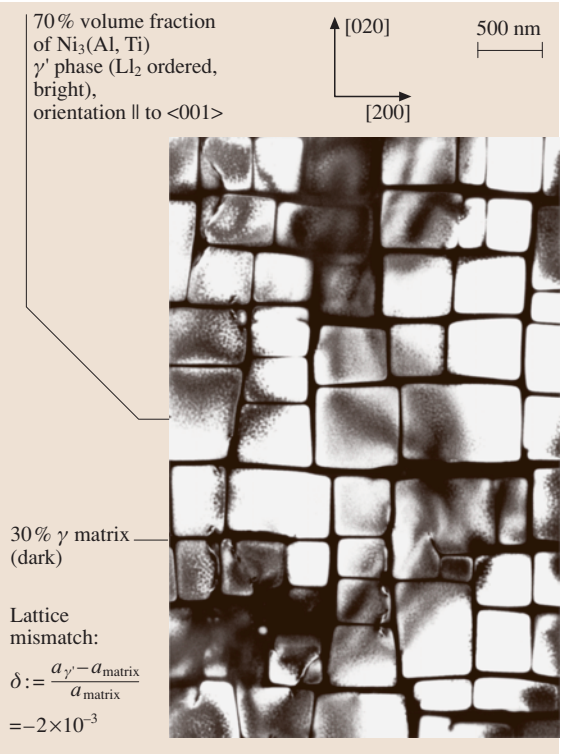


Fig. 3.153 Al–Ni phase diagram, the phases of interest are Ni solid solution γ and the $L1_2$ ordered coherent Ni3Al γ phase. Considerable interest has been given also to the B2 ordered NiAl β phase as a high-temperature structural material due to its very high melting point (after [3.1])



tuated recently with noble elements such as rhenium and ruthenium (Fig. 3.154). Finally, Cr is deliberately added in large concentrations of > 10 wt. %, typically around 20 wt. %, for chromia scale formation and protection against oxidation up to about 1000 °C.

A compilation of the most commonly employed wrought and cast Ni-based superalloys and their chemical composition can be found in [3.1]. Data for mechanical properties as a function of temperature are also extensively tabulated in [3.1], some characteristic examples for superalloys and their field of application are also listed in Table 3.38. Figure 3.155 illustrates some typical results of long-term stress rupture tests, demonstrating the suitability of Ni-based superalloys for applications in gas turbines and related power-generation applications.

3.7.6 Co and Its Alloys

General Properties

Due to their neighborhood in the periodic table, there are many analogies between Co and Ni. Like

Fig. 3.154 TEM micrograph of a second-generation Ni-based single-crystalline superalloy CMSX 4 (Courtesy of U. Glatzel, University of Bayreuth, Germany) ◀

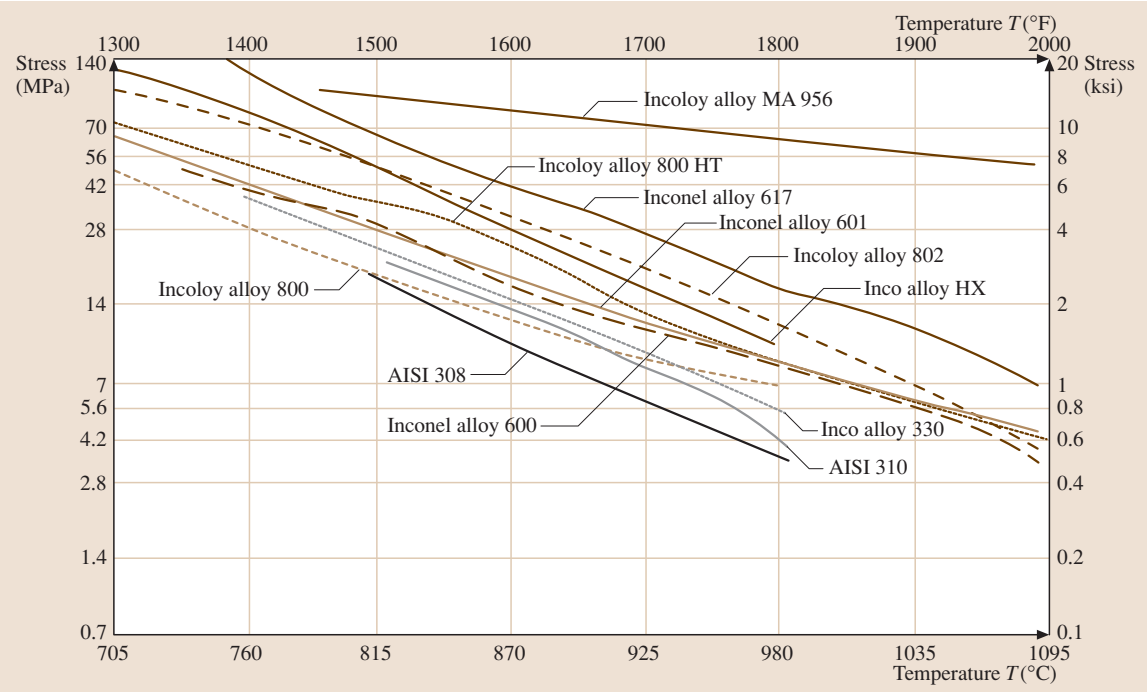


Fig. 3.155 Rupture strength (10 000 h) of Ni-based superalloys in comparison with selected stainless steels (after [3.1])

Ni cobalt also possesses physical properties which are similar to Fe [3.1]. However, at room temperature it exhibits a hexagonally closed-packed crystal structure like Mg, which undergoes an allotropic transformation at into an fcc crystal structure above approximately 660 K. Therefore, wrought deformation requires elevated temperatures and structural applications rely on the intrinsically high hardness of Co alloys, mainly manufactured via casting or powder-metallurgical technologies. Besides its use as an alloying element in steel, Co is frequently used as a component for many inorganic compounds such as a colorizer for glass and ceramics or in battery applications. While applications as surgical implant alloys and corrosion-resistant alloys are not treated further here, Co-based alloys may be grouped according to their field of application into the following three main categories.

Co-Based Hard-Facing Alloys

Co-based alloys with a carbon content in the range 1–3 wt. % C are widely used as wear-resistant hard-facing materials and weld overlays. In analogy to Ni-base alloys carbides of the type $M_{23}C_6$, M_6C and MC are formed depending on composition and heat treatment and *heavy* elements such as Mo, Ta, and W are deliberately added for solid-solution strengthening. Finally, a high Cr content of typically > 15 wt. % provides oxidation and hot corrosion resistance by chromia scale formation. In order to provide the necessary wear resistance under abrasive conditions, the microstructure of hard-facing Co-based alloys consists of a rather coarse dispersion of hard carbide phases embedded in a tough Co-rich metallic matrix. Due to the high C content of up to 3 wt. % the carbide volume fraction can exceed 50%. As a consequence of this, hot hardness values can exceed 500 HV at 650 °C (1200°F) and compressive strength can approach 2000 MPa trading off, however, for tensile ductility (< 1%) and UTS (around 800 MPa). Among some others, the most commonly known family of hard-facing Co-based alloys is designated *stellites*. For further details see [3.1] and Chap. 5 on tribology.

Co-Based Superalloys

Both wrought and cast Co-based superalloys differ significantly in chemical composition from their hard-facing counterparts as follows. First, they are based on the high-temperature face-centered cubic crystal structure, which is stabilized between RT and the melting

point by alloying with > 10 wt. % Ni. Second, for enabling wrought deformation the carbon content is reduced to values below 0.5 wt. %, which forms fine and homogeneously distributed carbides for dispersion strengthening. Finally, like for many Ni-based superalloys, in some Co-based superalloys the addition of Al and Ti serves to form the coherent ordered $Co_3(Al,Ti)$ phase, thus leading to precipitation strengthening by age hardening. These (often investment-cast) alloys are used in the very hot parts of gas turbines because of their excellent oxidation resistance. By contrast, they are inferior to Ni-based superalloys regarding creep strength (Fig. 3.152). A survey of the temperature dependence of the 1000 h rupture stress of typical Co-based superalloys is displayed in Fig. 3.156.

Cemented Carbides

Cemented carbides, also called *hardmetals*, can be considered as powder-metallurgically manufactured composite materials consisting of (rather coarse) carbide particulates embedded in a metallic Co-based matrix (*binder*). Most commonly, tungsten carbide (WC) is used for the particulates while the elements Ta, Nb, and Ti are deliberately added for economical reasons and to form complex multigrade cemented carbides. Then, the term *cermets* is used occasionally. Cobalt is the element of choice for the binder since it wets the carbides particularly well. Usually, Ni is added to the binder phase to increase corrosion and oxidation resistance. The main field of application is as grinding and turning tools for *difficult-to-machine* materials. Table 3.39 lists selected relevant hardmetals and their main properties.

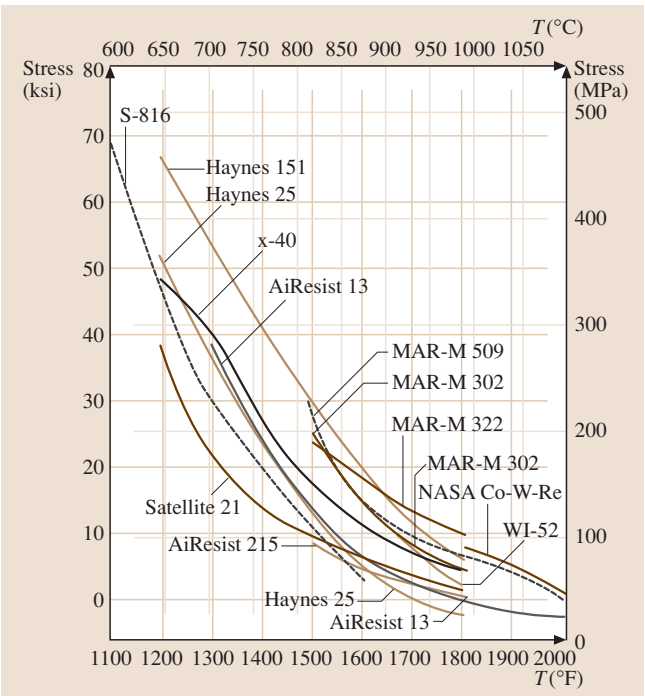


Fig. 3.156 Temperature dependence of 1000 h rupture stress of cast Co-based superalloys (after [3.1])

3.7.7 Copper and Its Alloys

The most striking evidence of the presence of copper in constructions is that as rooftops, where it is distinctively marked (after 10–15 years) by a green-colored layer of copper acetate, which prevents further

Table 3.40 Designation of Cu and its alloys (according to UNS)

Wrought alloys	
C100xx–C159xx	Commercially pure Cu
C160xx–C199xx	Nearly pure Cu, age hardenable
C2xxxx	Cu–Zn (classical brass)
C3xxxx	Cu–Zn–Pb (leaded brass)
C4xxxx	Cu–Zn–Sn (tin bronze)
C5xxxx	Cu–Sn (classical bronze) and Cu–Sn–Pb (phosphor bronze)
C6xxxx	Cu–Al (aluminum bronze), Cu–Si (silicon bronze), Cu-Zn-Mn (magnaese bronze)
C7xxxx	Cu–Ni (cupronickel), Cu–Ni–Zn (nickel silver)
Cast alloys	
C800xx–C811xx	Commercially pure Cu
C813xx–C828xx	95–99% Cu
C833xx–C899xx	Cu–Zn alloys containing Sn, Pb, Mn, or Si
C9xxxx	Other Cu alloys, including tin bronze, aluminum bronze, cupronickel, and nickel silver

Table 3.41 Composition and properties of characteristic unalloyed coppers (after [3.1])

Material	UNS no.	Purity; other elements (wt.%)	Yield stress $R_{po.2}$ (MPa)	Ultimate tensile strength R_m (MPa)	Fracture strain A_f (%)	Thermal conductivity κ (W m^{-1} K^{-1})	Electrical resistivity ρ ($\mu\,\Omega$ cm)
Pure Cu (oxygen-free electronic)	C10100	99.00 Cu	69–365	221–455	4–55	392	1.741
Pure Cu (oxygen free)	C10200	99.95 Cu	69–365	221–455	4–55	397	1.741
Electrolytic tough pitch Cu	C11000	99.90 Cu – 0.04 O	69–365	224–455	4–55	397	1.707
Oxygen-free low phosphorus Cu	C10800	99.95 Cu – 0.009 P	69–345	221–379	4–50	397	2.028
Phosphorus deoxidized arsenical Cu	C14200	99.68 Cu – 0.35 As – 0.02 P	69–345	221–379	8–45	397	3.831

corrosion (the statue of liberty is referred to as a prominent example). Copper and copper alloys have been in use for about 11 000 years, with Cu–Sn (bronze) probably being the first alloy of all. They generally have good corrosion resistance, excellent electrical and thermal conductivity, and their fabrication is easy due to the excellent formability (ductility). The favorable combination of electrical, mechanical, and corrosion properties aided the establishment of Cu as a structural material. On the other hand, Cu is susceptible to hydrogen embrittlement and stress corrosion cracking and has a relatively low strength-to-weight ratio. Comprehensive treatments and data of copper and copper alloys are given in [3.152–154]. The designation system of Cu alloys is given in Table 3.40.

Pure Copper

With its high electrical conductivity pure copper is primarily used for cables, wires, electrical contacts, and other electrical devices. A conductivity of 100% IACS

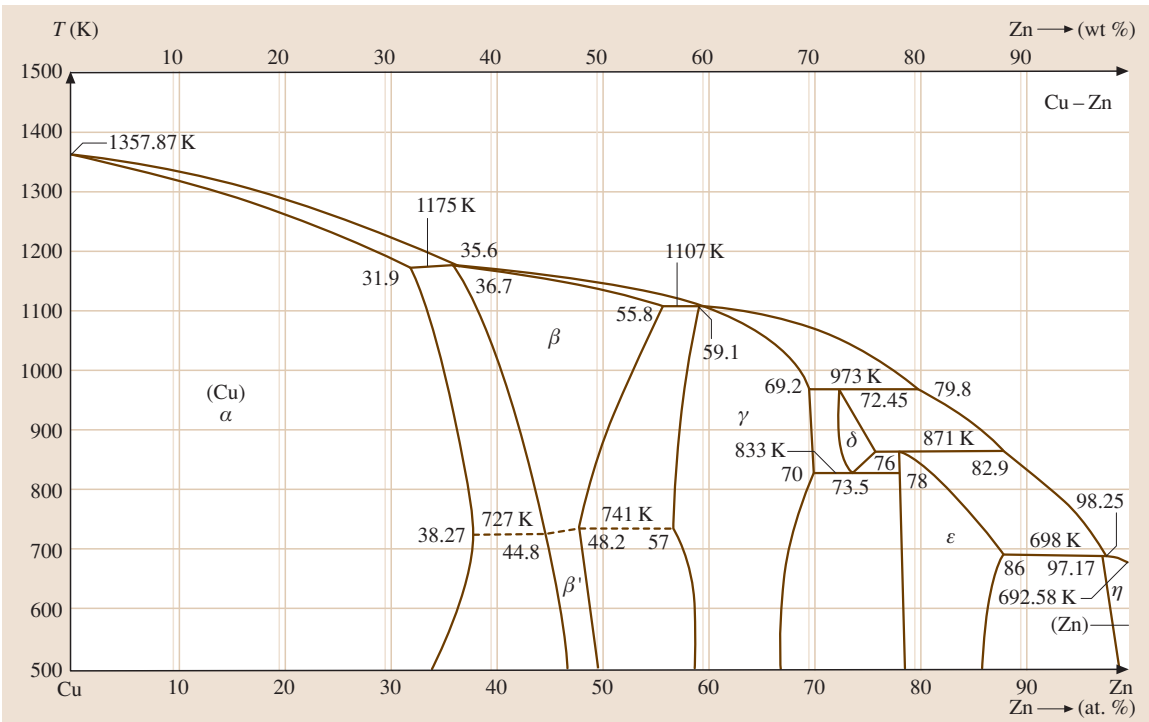


Fig. 3.157 Cu–Zn phase diagram (after [3.1])

Table 3.42 Designation, composition, and natural color of some brasses

Brass designation	Zn content (%)	Color
Gilding metal	5	Copper red
Commercial bronze	10	Golden
Red brass	15	Red gold
Yellow brass	35	Yellow
Muntz metal ($\alpha+\beta$)	40	Yellow gold

(International Annealed Copper Standards) corresponds to a resistivity of 1.72438 μ Ω cm. However, the properties of Cu are subject to dramatic changes with varying alloy content, i. e., the conductivity decreases substantially with increasing impurity content. Small oxygen additions of up to about 0.04% (electrolytic tough pitch copper) can bind metallic impurities to form oxides and therefore lead to an increase of the conductivity (Table 3.41), on the one hand. On the other hand, the presence of oxygen in Cu diminishes weldability, since hydrogen diffuses into the metal and interacts with oxide to form steam, which leads to cracking. For torch welding and brazing copper must be deoxidized, for example, by the addition of a small amount of phosphorus, which, however, lowers the electrical conductivity substantially but allows the material to be used in plumbing devices.

Copper Alloys

Elements which are solid-solution strengtheners in copper include Zn, Sn, Al, and Si, whereas Be, Cd, Zr, and Cr are suitable for age hardening. *Age-hardenable* alloys with small amounts of alloying additions (up to about 3%) can reach very high strength levels (yield stress $R_{pO.2} > 1300$ MPa at RT in the case of copper beryllium), offer high stiffness, and are nonsparking.

The term *brass* has been established for binary Cu−Zn alloys (Fig. 3.157) but is nowadays used for alloys containing additional components such as Pb, Fe, Ni, Al, and Si as well. Brasses are less expensive than pure Cu and can have different microstructures which depend on Zn content. Pure α-(Cu) solid solutions (up to about 38% Zn) are cold-working alloys. On increasing Zn content the natural color of brass changes form copper-like red (5% Zn) to yellow–gold (40% Zn) (Table 3.42). The *Muntz metal brass* is a binary $\alpha+\beta$ alloy with high strength and still reasonable ductility.

The most important properties of selected commonly used brasses are summarized in Table 3.43.

Wrought products of brasses and bronzes are used in automobile radiators, heat exchangers, and home heating systems, as pipes, valves, and fittings in carrying potable water and as springs, fasteners, hardware, small gears, and cams, to give a few examples. Cast leaded red and semi-red brasses find their application as lower-pressure-rating valves, fitting, and pump components as well as commercial plumbing fixtures, cocks, faucets, and certain lower-pressure valves. General hardware, ornamental parts, parts in contact with hydrocarbon fuels, and plumbing fixtures are made from yellow leaded brass, and high-strength (manganese-containing) yellow brass is suitable for structural, heavy-duty bearings,

Table 3.43 Composition and properties of characteristic brasses, bronzes, Cu−Ni and Cu−Ni−Zn alloys (after [3.1])

Material	UNS no.	Composition	Yield strength (MPa)	Tensile strength (MPa)	Elongation (%)	Thermal conductivity κ (W m^{-1} K^{-1})	Electrical resistivity ρ (μ Ω cm)
Gilding metal (cap copper)	C21000	95Cu–5Zn	69–400	234–441	8–45	234	3.079
Red brass	C23000	85Cu–15Zn	69–434	269–724	3–55	159	3.918
Yellow brass	C26800	65Cu–35Zn	97–427	317–883	3–65	121	6.631
Muntz metal	C28000	60Cu–40Zn	145–379	372–510	10–52	126	6.157
Free-cutting brass	C36000	61.5Cu–35.5Zn–3Pb	124–310	338–469	18–53	109	6.631
High-tensile brass (architecture bronze)	C38500	57Cu–40Zn–3Pb	138	414	30	88–109	8.620
Aluminum bronze	C60800	95Cu–5Al	186	414	55	85	9.741
Aluminum bronze	C63000	Cu–9.5Al–4Fe–5Ni–1Mn	345–517	621–814	15–20	62	13.26
Phosphor bronze D	C52400	90Cu–10Sn	193	455–1014	3–70	63	12.32
Silicon bronze A	C65500	97Cu–3Si	145–483	386–1000	3–63	50	21.29
Copper nickel	C71500	67Cu–31Ni–0.7Fe–0.5Be	138–483	372–517	15–45	21	38.31
Nickel silver 10%	C74500	65Cu–25Zn–10Ni	124–524	338–896	1–50	37	20.75

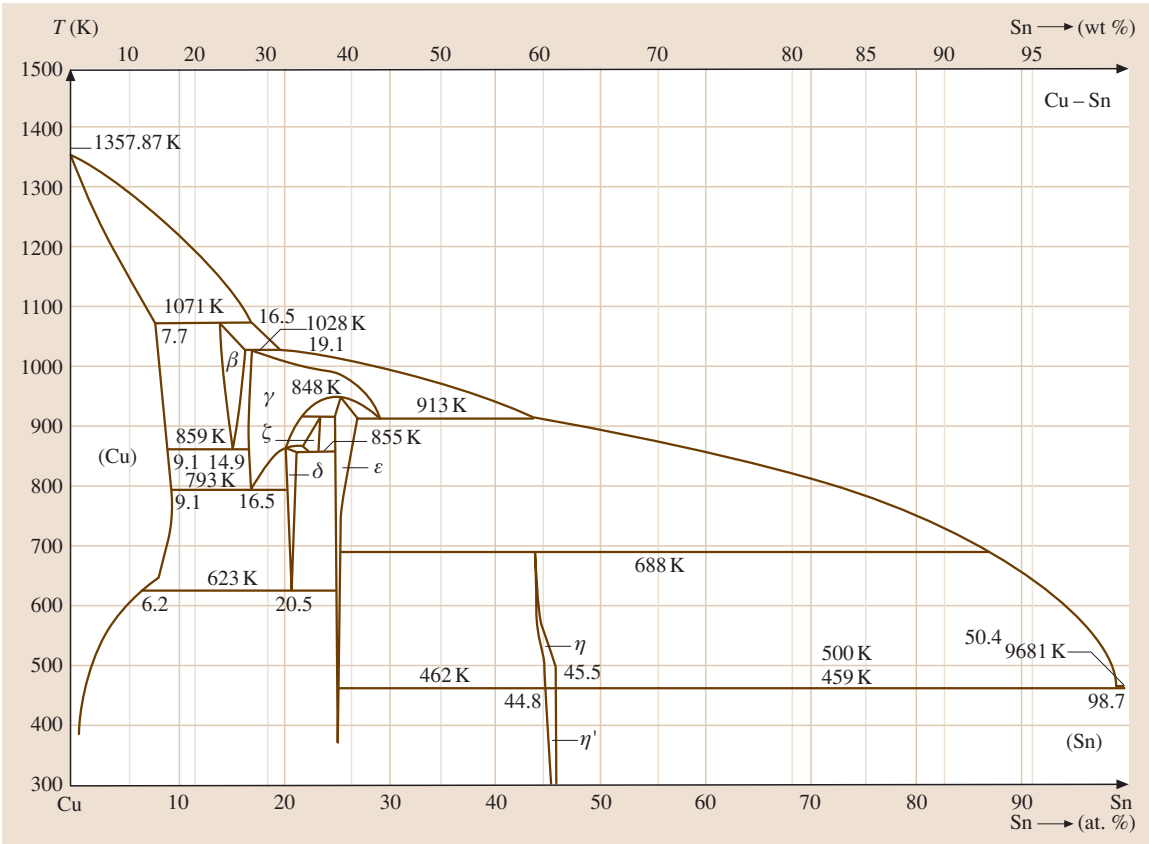


Fig. 3.158 Cu−Sn phase diagram (after [3.1])

hold-down nuts, gears, valve stems, and some marine fittings.

Bronzes are Cu−Sn- (Fig. 3.158), Cu−Al-, and Cu−Si-based alloys. Tin and aluminum are the most effective solid-solution strengtheners in copper. Cast products of tin bronzes are used as high-quality valves, fittings, and pressure vessel for applications at temperatures of up to 290 °C, special bearings, pump parts, gears, and steam fittings. Aluminum and silicon bronzes have very good strength, excellent formability, and good toughness. They are used as gears, slides gibs, cams, bushings, bearings, molds, forming dies, combustion engine components, valve stems, spark-resistant tools, and in marine applications such as propellers, impellers, and hydrofoils. The properties of some broadly used bronzes are given in Table 3.43.

Copper–nickel alloys show excellent corrosion resistance against seawater. Accordingly, they are used in shipboard components, power plants in costal areas, and saline-water conversion installations. Since Ni in Cu leads to a drastic decrease in electrical and thermal conductivity Cu−Ni alloys are also suitable for cryogenic applications.

3.7.8 Polymers

Polymers (polymer materials, polymeric materials, solid polymers, macromolecular materials) consist of very large molecules (chain molecules, macro molecules) which are synthesized from small molecules (monomers, monomer units) by a chemical reaction called polymerization (polyethylene, polyvinylchloride, polyurethane) or they are modified natural products (modified silk, regenerated cellulose) [3.155, 156].

The polymerization reactions can be classified into four groups [3.157]. *Chain polymerization* proceeds by the reaction of a monomer unit with the reactive site at the end of a polymer chain. These are mostly reactions via a radical mechanism [3.158]. The terminus *condensation chain polymerization* is used in

Table 3.44 Examples of widely used polymer materials and their abbreviations, characteristic backbone units, element groups within the backbone, and trademarks

Polymer	Backbone unit	Backbone	Trademarks
Organic polymers			
Polyethylene (PE)	$-CH_2-CH_2-$	−C−C−C−C−	Polythen, Lupolen, Hostalen
Polypropylene (PP)	$-CH_2-(CH_3)-CH_2-$	−C−C−C−C−	Hostalen, PPH, Luparen
Polyvinylchloride (PVC)	$-CH_2-CHCl-$	−C−C−C−C−	Hostalit, Vinidur, Vinylite
Polystyrene (PS)	$-CH_2-CH(C_6H_5)-$	−C−C−C−C−	Styroflex, Vestyron, Styropor (foam)
Polytetrafluorethylene (PTFE)	$-CF_2-CF_2-$	−C−C−C−C−	Teflon, Hostaflon
Polyamide (PA)	$-(CH_2)_6-NH-CO-(CH_2)_6-$	−C−N−C−C−	Nylon, Perlon
Polyethylene terephthalate (PET)	$-O-CO-C_6H_4-CO-O-CH_2-CH_2-$	−C−O−C−C−C−	Trevira (fiber), Diolen, Mylar (folie)
Polyurethan (PUR)	−NH−CO−O−	−C−C−N−C−O−C−C	
Polycarbonate (PC)	−O−CO−O−R	−C−O−C−C−	
Polyphenylene sulfide (PPS)	$-C_6H_4-S-$	−C−S−C−	Noxon, Ryton, Sulfar (fiber)
Inorganic polymers			
Polyphosphazene	$-N=PCl_2-$	−N=P−	
Polysiloxane (polydimethylsiloxane)	$O-Si(CH_3)_2-O-$	−Si−O−Si−O−	
Polysilane		−Si−Si−Si−Si−	

cases where a low-molar-mass byproduct is formed during polymerization. In *polycondensation* already generated polymer chains react with each other or with a monomer unit whereby a low-molar-mass byproduct is generated, for example, water as a byproduct in the reaction of an −OH group (alcohol group) with a −COOH group (organic acid group) resulting in an ester group. During *polyaddition*, growth of the polymer chains proceeds by an addition reaction between molecules of all degrees of polymerization or monomer units.

The annual world production of polymer materials is about 150–200 Mt. Some polymer materials are produced in amounts of more than 1 Mt/year (polypropylene about 14 Mt/year, which is about the same amount as for cotton), whereas others are polymer materials for special purposes with only small production volumes. Beside the use of bulk polymers as engineering materials a great amount of polymers is fabricated in the shape of fibers for manufacturing fabric, packaging films, paintings, thermal isolation materials (foam), and, for example, artificial leather.

Chemical Composition and Molecular Structure

For the presentation of polymer molecules the monomer unit is enclosed in brackets [ ] and an index (n) shows that a certain number of monomer units react to form the backbone of the polymer molecules. The polymerization of ethylene to polyethylene, for example, is written as $n CH_2{=}CH_2 \rightarrow [-CH_2-CH_2-]_n$, where the last part represents the whole molecule $CH_3-CH_2-CH_2...CH_2-CH_2-CH_3$, with n being between some hundreds and some millions.

Most of the polymers which are used as engineering materials are organic polymers with backbones (main chains) consisting of C−C bonds, or they contain bondings between C and other chemical elements (Table 3.44). Polymers with a backbone containing no carbon atoms are regarded as inorganic polymers. For most polymers common abbreviations are used and trademarks exist (Table 3.44).

Polymer materials can be classified, e.g., by their specific molecular structure and the resulting mechanical properties at different temperatures into *thermoplastics*, *elastomers*, and *duromers* [3.159].

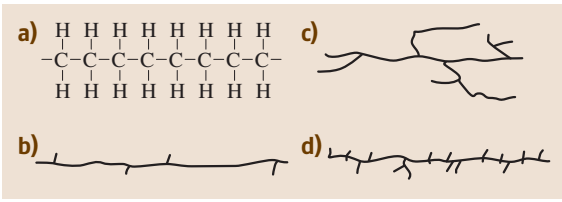


Fig. 3.159a–d Examples of linear polymer molecules: **(a)** theoretical backbone with carbon–carbon bonds; no side chains, **(b)** backbone with only a few small side chains, ≈ 10 side chains/1000 C atoms, example: high-density PE; **(c)** backbone with longer side chains/branches, example: low-density PE; **(d)** a great number of side chains attached to the backbone, example very low-density PE

Thermoplastics. *Thermoplastics* show good strength and high Young's modulus at RT and they are plastically deformable at elevated temperatures, in most cases above 100 °C. They consist in their simplest molecule structure of linear molecules with no branches (Fig. 3.159). In technical products small (e.g., $-CH_3$ groups) or larger side chains (short −C−C− chains) are attached to the main chain, forming a branched polymer. The degree of branching determines the density of solid polymers, because with increasing branching the possibility of a dense arrangement of the macromolecules decreases.

A typical example is polyethylene, with a density of 0.91–0.94 g/cm^3 for the strong branched low-density PE (LDPE) and a density of 0.94–0.97 g/cm^3 for the weakly branched high-density PE (HDPE). Regarding thermoplastics, within chain molecules there exist very strong intramolecular covalent bondings (bonding energy of the −C−C-bonding: 348 kJ/mol), whereas between neighboring molecule chains only weak intermolecular bonds with small bonding energies are present (Van der Waals bond: 0.5–5 kJ/mol, hydrogen bond: ≈ 7 kJ/mol). Therefore the chain molecules can, already around room temperature (rubber-like polymers, elastomers) or at elevated temperatures (thermoplastics), shifted with respect to each other, and such polymer solids can be deformed elastically or plastically. The molecular structure of thermoplastics can be distinguished by the kind of atoms building the backbone and by the kind of atoms or chemical groups attached to the backbone (Table 3.45). The side groups determine the polymer properties to a large extent, because they influence the strength of the intermolecular bonding.

Another significant parameter that determines the properties of polymer solids results from the mean size of the macromolecules (degree of polymerization, mean chain length, mean molar mass), and, because the polymer molecules show no unit length, the deviation of the molecule size, which depends on the production parameters.

Elastomers. *Elastomers* (rubber-like polymers) consist, similarly to thermoplastics, of linear molecules, but the molecule chains are bridged by small-molecule segments via covalent bondings. The molecules can therefore undergo a strong elastic deformation at *room temperature*. This effect is due to the stretching of the molecules out of the disordered state if a load is applied, and a re-deformation into the random tangle of molecules due to the increased entropy, after the load is released.

Duromers. *Duromers* consist of a three-dimensional molecule network, bridged by covalent bondings. Even at elevated temperatures they undergo no plastic deformation and can, in most cases, be heated up to their decomposition temperature without any elastic or plastic deformation. Most duromers are thermosets (phenolics, unsaturated polyesters, epoxy resins, and polyurethanes) which solidify by an exothermal chemical reaction (curing). Thermosets are obtained by moulding a thermoplastic material into the desired shape, which is then cross-linked. The curing reaction can be initiated at room temperature (RT) by mixing the components, or it starts at an elevated temperature, or irradiation by energetic radiation (ultraviolet light, laser beam, or electron beam) is applied.

Table 3.45 Examples of chemical groups/atoms on the backbone of linear polymers

	X	*Y*	**Polymer**
$-CH_2-CH_2-CH_2-CH_2-CH_2-$	H	H	Polyethylene
	CH_3	H	Polypropylene
	Cl	H	Polyvinylchloride
	C_6H_5	H	Polystyrene
	CH_3	$COOCH_3$	Polymethylmethacrylate

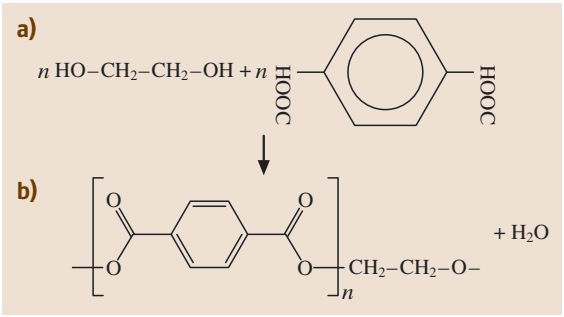


Fig. 3.160 PET formation by polycondensation of ethylene glycol with terephthalic acid

Most polymers are formed by chain polymerization from one type of monomer (PE, PP, PVC, PS). For example, PE is obtained by polymerization of ethylene: $n(CH_2{=}CH_2) \rightarrow -[-CH_2-CH_2-]_n$. Another possibility is that two different monomers containing different types of chemical groups react with each other (PA, PC, PET, polyurethane), thus forming a polymer unit built by two molecules; for example, PET is obtained by polycondensation of ethylene glycol with terephthalic acid whereby water is generated as a byproduct, as exemplified in Fig. 3.160.

In copolymers the polymer chain is composed of two or more types of monomers. The monomers can be arranged randomly, alternating, or as blocks (short molecule chains, consisting of the same monomer units). Another version is that blocks built from one type of monomer are fixed as side chains onto a backbone built from another monomer type. Common copolymers are polystyrene-butadiene-rubber (SBR) and acrylonitrile-butadiene-styrene (ABS).

The properties of polymeric materials can be further tailored by mixing or blending two or more polymers [3.160]. One goal of blending is to obtain materials with greater impact toughness than the pure polymers, whereby one component functions as a toughener for the other. In high-impact polystyrene, the high modulus of polystyrene is combined with the high impact strength of rubber particles (polybutadiene). Other examples for blends are PP-PC, PVC-ABS, and PE-PTFE.

The toughness of otherwise stiff polymers and the glass-transition temperature (see below) are increased by mixing low-molar-mass substances (plasticizers) into the polymers. The most common case is dioctyl phthalate as a plasticizer for PVC. A similar effect results if up to 8% water is incorporated into polyamide.

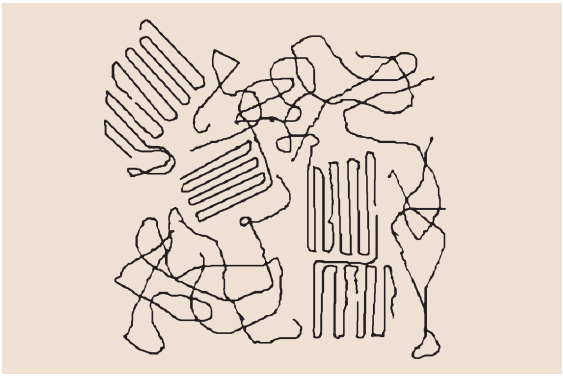


Fig. 3.161 Amorphous and crystalline regions in the morphology of a semicrystalline polymer

In a process known as cross-linking molecules are linked to one another to increase the temperature resistance, long-time creep strength, and insensitivity to stress cracks. Cross-linking can be achieved very precisely by irradiating plastics with high-energy electron beams or gamma rays. This optimization can be applied even to pure and widely used plastics, such as PE and PVC. Another approach is to add special compounds which are cross-linkable by irradiation, thus fixing two polymer chains to each other. The advantage of the cross-linking technique is that the properties are modified after the components have been formed into parts and that the process takes place at room temperature and at normal pressure.

Microstructure of Polymer Materials

Linear polymers can have a disordered state (amorphous) or a semicrystalline arrangement of the molecules (Fig. 3.161).

The amorphous state is characterized by a random tangle of molecules. In semicrystalline polymers some parts of the polymer molecules are ordered in the shape of folded-chain lamellae which are eventually arranged to form blocks. The degree of crystallinity can be estimated by density measurement, by thermal analysis/differential scanning calorimetry (DSC) or by

Table 3.46 Degree of crystallinity of common polymer materials

Polymer	Degrees of crystallinity (%)
Low-density PE	45–75
High-density PE	65–95
PP fiber	55–60
PET fiber	20–60

XRD. It depends on the number and length of the side chains on the molecule backbone (degree of branching) and determines the density and the elastic moduli of a polymer material. For polyethylene the degree of crystallinity varies from 45% for low-density PE to 95% for high-density PE (Table 3.46).

In some semicrystalline polymers the folded-chain lamellae grow, starting at a nucleus, outwards, yielding spherulites [3.161]. The size of the spherulites can be modified by the addition of nucleants. Toughness and light transparency decrease with increasing spherulite size. Spherulites are visible by polarized-light microscopy of microtome sections (about 25 µm thick) (Fig. 3.162).

The polymer molecules can be strongly oriented parallel to the flow direction during production processes such as injection moulding due to the high viscosity of the melt. This occurs especially if metallic parts are used as inserts by which the melt flow is divided and then reunited after the flow around the insert [3.162, 163]. As a result one recognizes regions which show strong anisotropic mechanical properties within the normally isotropic polymer solid [3.164]. The aligned molecules can result in substantial residual stresses and give rise to crack initiation even when a low external load is applied. This molecular alignment can be lowered by relaxation at elevated temperatures, whereby changes of the shape can occur. This effect has to be taken into account if parts made from polymer materials are heated during further manufacturing or while in use. An application of the relaxation effect is the use of polymer films for shrink packaging, where an article is wrapped at room temperature and the film shrinks upon heating.

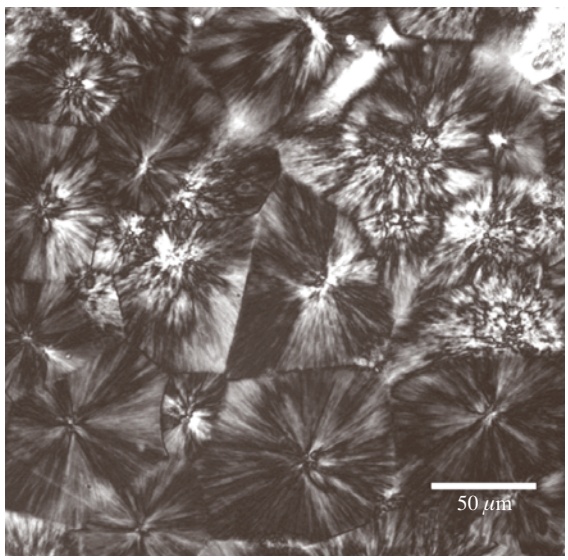


Fig. 3.162 Spherulites in polypropylene; transmission light microscopy of a microtome section; polarized light

Thermal Properties

The strength of duroplastic materials does not change much with increasing temperature. They do not melt at all due to the three-dimensional molecule network but rather start to decompose. During heating of thermoplastic and elastomeric polymers temperature ranges with different polymer properties are observed: strong, hard, and brittle at low temperatures but ductile and deformable at increased temperatures, and finally changing from a solid to the state of a viscous melt. The transitions between the different states which are due to the arrangement and the mobility of the molecules can be investigated by DSC [3.165] whereby the exothermic and endothermic heat flux is registered (Fig. 3.163).

At a material-specific temperature an endothermic hump appears, extending over a certain temperature range. This is the transition into the glassy state due to the increased mobility of the molecule segments in the amorphous parts of the microstructure and is accompanied by an enormous decrease of viscosity and, thus, strength. The glass-transition temperature (T_g), which can be defined as the temperature at the first inflection point of the graph, is heating rate dependent and is specific to different materials (Table 3.47). The high level of heat flux is maintained, because the specific heat of the glassy polymer is greater than that of the solid polymer.

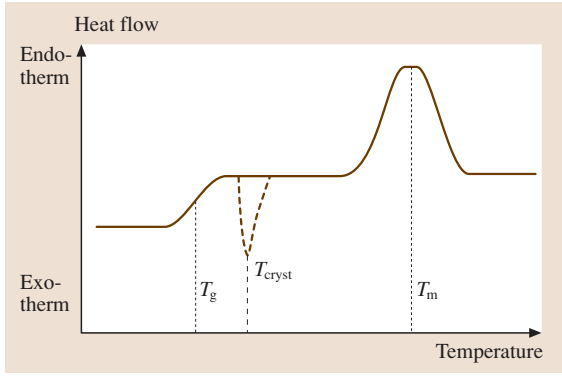


Fig. 3.163 DSC result of a partially crystalline polymer; schematic heating curve with characteristic transition points: glass transition T_g, crystallization temperature T_{cryst}, and melting point T_m

Table 3.47 Typical glass-transition temperatures (T_g) and melting temperature (T_m) of polymers

Polymer	T_g (°C)	T_m (°C)
Polyethylene (PE)	–120	130
Polypropylene (PP)	–15	170
Polystyrene (PS)	90	200
Polyvinylchloride, amorphous (PVC)	80	–
Polyvinylchloride, partially crystalline (PVC)	80	210
Polytrafluorethylene (PTFE)	–115	330
Polymethylmethacrylate (PMMA)	45	160
Polyamide 6 (PA6)	75	230
Polyethyleneterephthalate (PET)	75	280

In very few cases and at very low temperature partial crystallization of the polymer can follow the transition into the glassy state, connected with an exothermic heat flow (dashed line in Fig. 3.163). With increasing temperature the crystalline regions of semicrystalline polymers will melt, which is indicated by another endothermic event, which represents the melting temperature (T_m) (Table 3.47). The large peak width, as compared with metals, is related to the nonuniformity of the polymer molecules and the high degree of imperfection of the polymer crystals. After that temperature range the polymer behaves like a high-viscosity melt.

Usually, technical applications of polymer materials are restricted to temperatures below the glass transition (polystyrene, PMMA, PET). In some cases polymers are also used above that temperature (polyisoprene, polybutadiene, polyethylene); then some polymers exhibit rubber-like properties.

Polymeric materials start to decompose or to be oxidized in air if heated over a certain temperature. In some cases they burn (PE, PP, PS) with a characteristic flame color. The decomposition is a process of thermal cracking of the material and/or an oxidation, whereby sometimes a characteristic smell of the fume occurs when the flame is blown out (PE: like burning candle or wax; PA: like burned hair or horn; PVC: stinging, acidic; PS: fruity). In some cases chemicals which are dangerous for humans or can alter other materials are set free. The burning of PUR generates toxic hydrocyanic acid. Overheating or burning of PVC yields hydrochloric acid vapor (HCl) which leads to the corrosion of parts made from Cu or other metals and is toxic to humans.

Mechanical Properties of Polymer Materials

As stated above the three basic types of polymers – thermoplastics, elastics, and duromers – show very different properties with different dependencies on temperature, which can be a reason for selecting a certain polymer. Choosing a polymeric material for a specific application [3.167] can be based on the difference in mechanical properties, such as tensile strength, impact, elastic behavior, but also often other properties have to be considered, such as density, corrosion resistance, or formability. The ratio of density and Young's modulus (Fig. 3.164) can be a potential criterion for selecting a certain polymer material for a certain application. Alternatively, the ratio of Young's modulus and impact

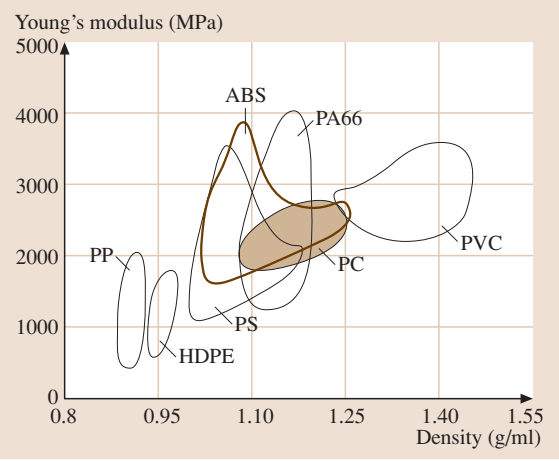


Fig. 3.164 Young's modulus versus density for selected thermoplastics polymers (after [3.166])

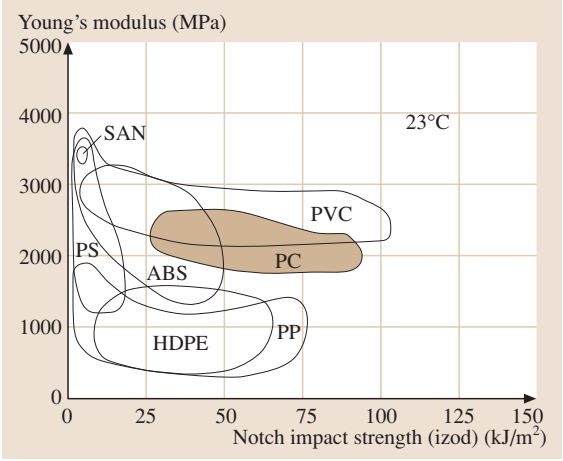


Fig. 3.165 Young's modulus versus notch impact strength for thermoplastics (after [3.166])

Table 3.48 Selected standard mechanical testing methods for polymer materials

Standard	Testing method
ISO 178:2001	Plastics – Determination of flexural properties
ISO 179-1:2000	Plastics – Determination of Charpy impact properties; Part 1: Noninstrumented impact test
ISO 179-2:1997	Plastics – Determination of Charpy impact properties; Part 2: Instrumented impact test
ISO 180:2000	Plastics – Determination of Izod impact strength
ISO 527-1:1993	Plastics – Determination of tensile properties; Part 1: General principles
ISO 527-2:1993	Plastics – Determination of tensile properties; Part 2: Test conditions for moulding and extrusion plastics
ISO 527-3:1995	Plastics – Determination of tensile properties; Part 3: Test conditions for films and sheets
ISO 6721-1:2001	Plastics – Determination of dynamic mechanical properties; Part 1: General principles
ISO 6721-2:1994	Plastics – Determination of dynamic mechanical properties; Part 2: Torsion-pendulum method
ISO 899-1:2003	Plastics – Determination of creep behavior; Part 1: Tensile creep
ISO 899-2:2003	Plastics – Determination of creep behavior; Part 2: Flexural creep by three-point loading
ISO 8256:2004	Plastics – Determination of tensile-impact strength
ISO 2039-1:2001	Plastics – Determination of hardness; Part 1: Ball indentation method
ISO 868:2003	Plastics and ebonite – Determination of indentation hardness by means of a durometer (Shore hardness)

strength of a polymer material may be taken into account (Fig. 3.165).

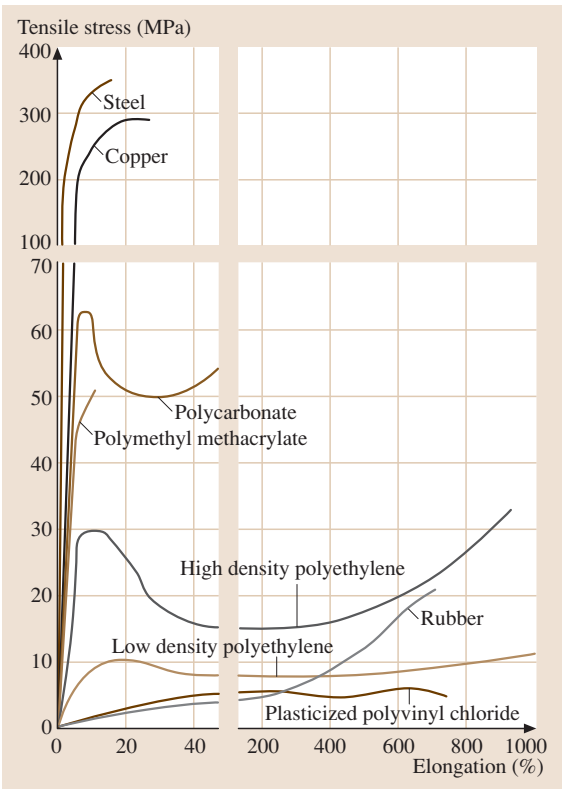


Fig. 3.166 Tensile stress–strain curves for polymers in comparison to copper and steel (after [3.166])

Standardized mechanical testing methods for polymer materials (Table 3.48) in most cases differ from those applied to other materials (Sect. 3.3). This is especially true for the size and shape of the specimens and the applied load [3.169].

Tensile stress–strain curves can be very different for polymer materials (Fig. 3.166) and they are strongly dependent on the testing temperature (Fig. 3.167). No linear region of the stress–strain curve of polymer materials exists from which the Young's modulus could be obtained. Therefore a *secant modulus* is calculated

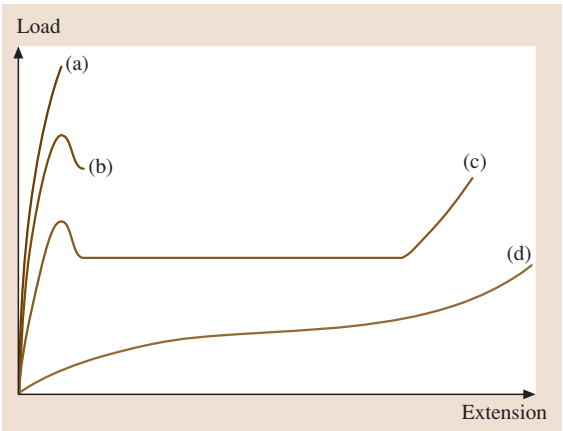


Fig. 3.167a–d Temperature dependence of the load–extension curve for a polymer; with increasing temperature: **(a)** brittle ductile, **(b)** homogeneous deformation, **(c)** necking and cold-drawing, **(d)** quasi-rubber-like behavior (after [3.168])

Table 3.49 Properties of common polymer materials

Material	Density (g/cm^3)	Young's Modulus (GPa)	Ball indentation hardness	Izod A at room temperature (kJ/m^2)
LDPE	0.92	0.2	50 *	2–35
HDPE	0.95	1	50 *	2–35
PP	0.9	1.5	70	3–10
PA 6,6	1.1	3	160	5–90
PVC	1.4	3	110	4; 40–70 **
PS	1.05	3.2	150	2–15
PC	1.2	2.5	110	80
ABS	1.05	3	95	10–35
PMMA	1.19	3.3	200	3
PTFE	2.1	0.75		16

* Shore D; ** with plasticizer

based on the slope of the stress–strain curve within a certain range of elongation, e.g., between 0.05% and 0.25% elongation

$$E_t = \frac{\sigma_{0.05} - \sigma_{0.25}}{\varepsilon_{0.05} - \varepsilon_{0.25}} \,. \qquad (3.94)$$

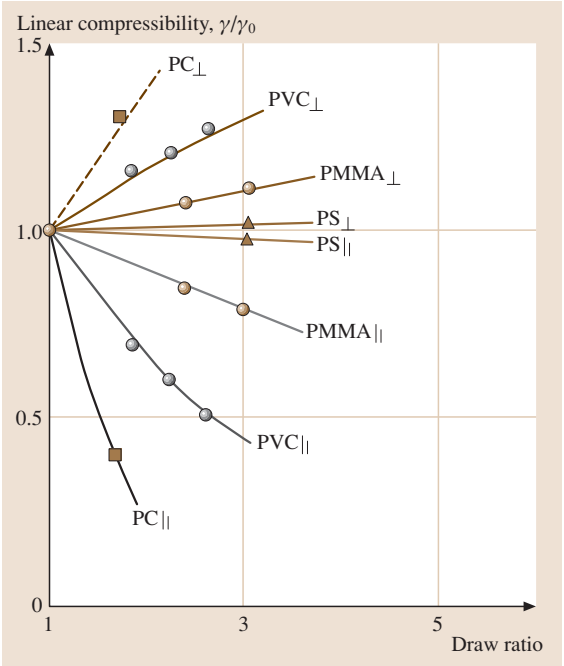


Fig. 3.168 Influence of the drawing ratio on the linear compressibility parallel and perpendicular to the drawing direction (after [3.166])

In general the values of the mechanical properties of polymer materials are inferior to those of metallic materials (Table 3.50).

The tensile strength generated by stretching the polymer chains during the manufacturing process can yield values of the ultimate tensile strength which are greater than those known for steel (for example, steel S355 ≈ 400 N/mm^2) (Table 3.51). An outstanding high tensile strength is exhibited by Aramid (poly-paraphenylene terephthalamide; Kevlar), which is used as a fiber.

The orientation of segments of the molecule chains as generated by the shaping process [3.170] (see above)

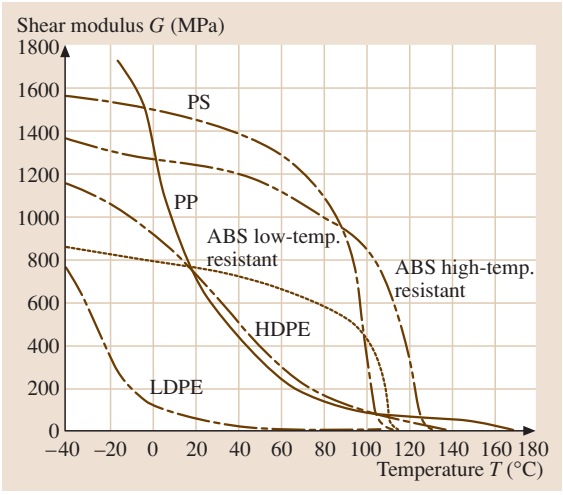


Fig. 3.169 Shear modulus versus temperature for several common polymer materials (after [3.166])

Table 3.50 Comparison of the specific ultimate tensile strength (tensile strength/density) with steel: value for Aramid set to 100

Material	Relative specific UTS
Aramid/KEVLAR	100
Glasfiber E	46
PA/nylon fiber	45
Low-carbon steel	19

has a significant influence on the mechanical properties (Fig. 3.168). For the determination of dynamic mechanical properties of polymers a torsion pendulum is used [3.171]. As a result the elastic shear modulus G and $\tan\delta$ are obtained. The shear modulus is strongly dependent on temperature (Fig. 3.169).

The mechanical properties of polymer materials can be further improved by fiber reinforcement [3.173, 174] (Sect. 3.7.10).

Polymer Interaction with Solvents

The dissolution of solid polymers in organic solvents or water starts with swelling, whereby the macromolecules are not degraded, which means that the chain length is not changed [3.175]. Only in some polymers are the chain molecules shortened by a chemical reaction with a chemical substance contained in a solvent. For example, the amid bondings in polyamides undergo hydrolysis under basic conditions (saponification), resulting in the generation of chain molecule fragments of different length. Swelling and subsequent dissolution are due to a competition of the intermolecular bonding forces between chains of the polymer, and the bonding forces between the macromolecules and the small solvent molecules, respectively. As a result, increasing numbers of solvent molecules penetrate the tangled polymer chain arrangement and lead to an increase of the volume of the polymer solid. This is accompanied by a lowering of the interaction forces between adjacent macromolecule segments and an increase of the mobility of the molecules with respect to each other and a loss of strength. The swelling and dissolution process may take up to several days or weeks at ambient temperature. Swelling often results in a sticky substance before the real dissolution happens. In some cases polymer solutions can be used as a glue which will have the strength of the starting polymer after the solvent has evaporated.

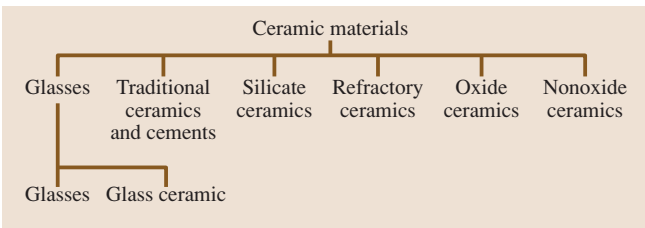


Fig. 3.170 Classification of ceramic materials on the basis of chemical composition (after [3.1])

Table 3.51 Solubility parameter for solvents and polymers [3.172]

Solvent	δ (MPa)$^{1/2}$	Polymer	δ (MPa)$^{1/2}$
n-hexane	14.9	Polyethylene	12.7
Benzene	18.8	Polystyrene	18.4

Some polymers can only incorporate a limited fraction of solvent into the solid. The interaction between a polymer and a selected solvent and therefore the solubility of the polymer can be predicted using the solubility parameter δ (Table 3.51), which is based on the cohesion forces, beside other factors [3.172]. As a rule, a substance can be regarded as a solvent if the difference of δ values is less than 2.

Aging and Corrosion

Aging of polymers is mainly due to chemical changes of the structure of the macromolecules accompanied by a shortening of the chain molecules, branching, cross-linking, and the generation of new chemical groups. A prerequisite for aging is the influence of light, especially UV light, and eventually oxygen from the air. As a result the polymer becomes brittle, cracks are generated, the quality of the surface is changed, and a loss of electrical insulation behavior will appear. Loss of plasticizer by diffusion also lowers the elasticity and the ductility, especially at lower temperatures. An especially dangerous situation is the interaction of a solvent or a solution and mechanical stress on a polymer part, leading to stress-corrosion failure.

3.7.9 Glass and Ceramics

Ceramic and glass materials are complex compounds and solid solutions containing metallic and nonmetallic elements, which are composed either by ionic or covalent bonds. Typical properties of glasses and ceramics include high hardness, high compressive strength, high brittleness, high melting point, and low electrical and thermal conductivity. There are several ways in which ceramics may be classified, such as by chemical composition, properties or applications. In Fig. 3.170 this

Table 3.52 Compositions and characteristics of some common commercial glasses (after [3.176])

Glas type	Composition (wt%)						Characteristics and applications
	SiO_2	Na_2O	CaO	Al_2O_3	B_2O_3	Other	
Fused silica	> 99.5						High melting temperature, very low coefficient of expansion (shock resistant)
96% Silica (Vycor)	96				4		Thermal shock and chemically resistant (laboratory ware)
Borosilicate (Pyrex)	81	3.5		2.5	13		Thermal shock and chemically resistant (ovenware)
Container (soda lime)	74	16	5	1		4MgO	Low melting temperature, easily worked, also durable
Fiberglass	55		16	15	10	4MgO	Easily drawn into fibers (glass-resin composites)
Optical flint	54	1				37PbO, $8K_2O$	High density and high index of refraction (optical lenses)
Glass-ceramic (Pyroceram)	43.5	14		30	5.5	$6.5TiO_2$, $0.5As_2O_3$	Easily fabricated; strong; resists thermal shock (ovenware)

classification is made on the basis of chemical composition [3.1].

In the following, a closer look at some of the ceramic materials listed in Fig. 3.170 will be made. Detailed treatments of ceramics are given in [3.177, 178].

Glasses

Glasses are solid materials which have become rigid without crystallization (amorphous structure, Sect. 3.1). The microstructure is based on SiO_4 tetrahedral units which possess short-range order and are connected to each other by bridging oxygen, resulting in a three-dimensional framework of strong Si−O−Si bonds. The main assets of glasses are their optical transparency, pronounced chemical resistance, high mechanical strength, and relatively low fabrication costs. Glasses usually contain other oxides, notably CaO, Na_2O, K_2O, and Al_2O_3, which influence the glass properties. Beside about 70% SiO_2 soda-lime glasses, which are used for windows and containers, additionally consist of Na_2O (soda) and CaO (lime). Further applications of glasses are as lenses (optical glasses), fiberglass, industrial and laboratory ware, and as metal-to-glass sealing and soldering. The compositions of some commercial glass materials are described in Table 3.52 [3.176].

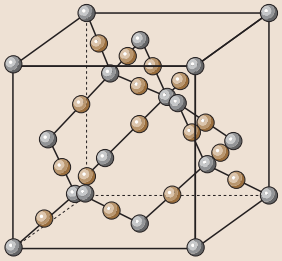
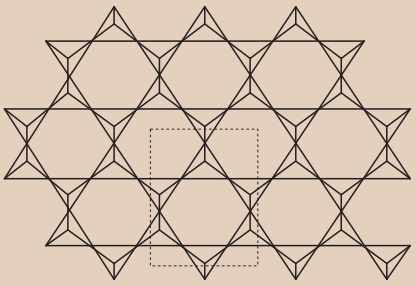
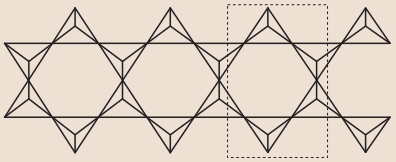
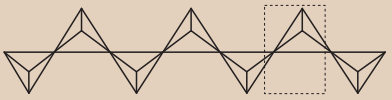
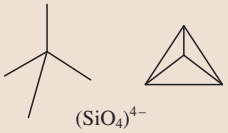
Glass Ceramics

Glass ceramics contain small amounts of nucleating agents (such as TiO_2 and ZrO_2) which induce crystallization of glasses when exposed to high temperatures. After melting and shaping of the glassy material, it is partly crystallized using a specific heat treatment at temperatures between 800 and 1200 °C. The residual glass phase occupies 5–50% of the volume and the crystalline phase has a grain size of 0.05–5 µm. In contrast to conventional ceramics, e.g., prepared by powder processing routes, glass ceramics are fully dense and pore-free, resulting in relatively high mechanical strength. Glass ceramics of the system $Li_2O-Al_2O_3-SiO_2$ show near-zero linear thermal expansion, such that the glass ceramic ware will not experience thermal shock. These materials also have a relatively high thermal conductivity and show exceptionally high dimensional and shape stability, even when subjected to considerable temperature variations. Glass ceramics are used in astronomical telescopes, as mirror spacers in lasers, as ovenware and tableware, as electrical insulators, and are utilized for architectural cladding, and for heat exchangers and regenerators.

Silicate Ceramics

Silicates are the most important constituents of the Earth's crust. Their structure, which is based on SiO_4 tetrahedrons (glasses are a derivative of silicates) depends on the actual composition. A three-dimensional network (quartz) is only stable when the ratio of O/Si is exactly 2. The addition of alkali or alkali-metal oxides to silica increases the overall O/Si ratio of the silicate and results in the progressive breakdown of the silicate structure into smaller units. In Table 3.53 the relationship of the O/Si ratio and the

Table 3.53 Relationship between silicate structure and the O/Si ratio

Structure	O/Si ratio	No. of oxygens per Si		Structure and examples
		Bridging	Nonbridging	
	2.00	4.0	0.0	Three-dimensional network quartz, tridymite, cristabolite are all polymorphs of silica
 Repeat unit $(Si_4O_{10})^{4-}$	2.50	3.0	1.0	Infinite sheets $Na_2Si_2O_5$ clays (kaolinite)
 Repeat unit $(Si_4O_{11})^{6-}$	2.75	2.5	1.5	Double chains, e.g., asbestos
 Repeat unit $(SiO_3)^{2-}$	3.00	2.0	2.0	Chains $(SiO_3)_n^{2n-}$, Na_2SiO_3, $MgSiO_3$
 $(SiO_4)^{4-}$	4.00	0.0	4.0	Isolated SiO_4^{4-}, tetrahedra Mg_2SiO_4 olivine, Li_4SiO_4

The simplest way to determine the number of nonbridging oxygens per Si is to divide the charge on the repeat unit by the number of Si atoms in the repeat unit

silicate structure is demonstrated. Silicates are particularly useful for electrotechnical, electronic, and high-temperature applications as well as in the processing of materials.

Refractory Ceramics

Refractory ceramics are another important group of ceramics that are utilized in large tonnages. These materials must withstand high stresses at high tem-

Table 3.54 Properties of fired refractory brick materials (after [3.1])

Brick (major chemical components)	Density ϱ (kg/m^3)	Melting temperature (°C)	Thermal conductivity κ (W/(m K))
Building brick	1842	1600	0.72
Chrome-magnesite brick (52 wt. % MgO, 23 wt% Cr_2O_3)	3100	3045	3.5
Fireclay brick (54 wt. % SiO_2, 40 wt% Al_2O_3)	2146–2243	1740	0.3–1.0
High-alumina brick (90–99 wt. % Al_2O_3)	2810–2970	1760–2030	3.12
Silica brick (95–99 wt. % SiO_2	1842	1765	1.5
Silicon carbide brick (80–90 wt. % SiC)	2595	2305	20.5
Zirconia (stabilized) brick	3925	2650	2.0

Table 3.55 Properties of commercial oxides according to DIN EN 60672 [3.1]

Oxide	MgO (C 820; 30% porosity)	Al_2O_3(> 99.9)	TiO_2 (C 310)	Beryllium oxide C 810	Partially stabilized ZrO_2
Density ρ (g/cm^3)	2.5	3.97–3.99	3.5	2.8	5–6
Young's modulus (GPa)	90	366–410	–	300	200–210
Bending strength (MPa)	50	550–600	70	150	500–1000
Coefficient of thermal expansion (RT) (10^{-6} K^{-1})	11–13	6.5–8.9	6–8	7–8.5	10–12.5
Thermal conductivity (RT) (W m^{-1} K^{-1})	6–10	38.9	3–4	150–220	1.5–3
Application examples	For insulation in sheathed thermo-couples; in resistive heating elements	In insulators; in electrotechnical equipment; as wear-resistant machine parts; in medical implants	In powder form as a pigment and filler material; in optical and catalytic applications	In heat sinks for electronic components	As thermal barrier coating of turbine blades

peratures without melting or decomposing and must remain nonreactive and inert when exposed to severe environments. Refractory ceramics are composed of coarse oxide particles bonded by a finer refractory material. The finer material usually melts during firing and bonds the remaining material. Refractory ceramics generally contain 20–25% porosity as an important microstructural variable that must be well controlled during manufacturing. They are used for various applications ranging from low- to intermediate-temperature building bricks to high-temperature applications, where magnesite, silicon carbide, and stabilized zirconia (also used as thermal barrier coatings of nickel-based turbine components) are suitable. Typical applications include furnace linings for metal refining, glass manufacturing, metallurgical heat treatment, and power generation. Depending on their chemical composition and reaction oxide refractories can be classified into acidic, basic, and neutral refractories. Fireclays are acidic refractories and are formable with the addition of water (castable and cements). Very high melting points are provided by chromite and chromite–magnesite ceramics, which are neutral refractories. Examples of commercial refractories are given in Table 3.54.

Oxide Ceramics

Oxide ceramics are treated as a separate group of ceramics in [3.1] since they are the most common constituents

Table 3.56 Properties and applications of advanced ceramics

Property	Application (examples)
Thermal	
Insulation	High-temperature furnace linings for insulation (oxide fibers such as silica, alumina, and zirconia)
Refractoriness	High-temperature furnace linings for insulation and containment of molten metals and slags
Thermal conductivity	Heat sinks for electronic packages (AlN)
Electrical and dielectric	
Conductivity	Heat elements for furnaces (SiC, ZrO_2, $MoSi_2$)
Ferroelectricity	Capacitors (Ba-titanate-based materials)
Low-voltage insulators	Ceramic insulation (porcelain, steatite, forsterite)
Insulators in electronic applications	Substrate for electronic packaging and electical insulators in general (Al_2O_3, AlN)
Insulators in hostile environments	Spark plugs (Al_2O_3)
Ion-conducting	Sensors, fuel cells, and solid electrolytes (ZrO_2, β-alumina, etc.)
Semiconducting	Thermistors and heating elements (oxides of Fe, Co, Mn)
Nonlineal I–V characteristics	Current surge protectors (Bi-doped ZnO, SiC)
Gas-sensitive conductivity	Gas sensors (SnO_2, ZnO)
Magnetic and superconductive	
Hard magnets	Ferrite magnets [(Ba, Sr)O × $6Fe_2O_3$]
Soft magnets	Transformer cores [(Zn, M)Fe_2O_3, with M = Mn, Co, Mg]; magnetic tapes (rare-earth garnets)
Superconductivity	Wires and SQUID magnetometers ($YBa_2Cu_3O_7$)
Optical	
Transparency	Windows (soda-lime glasses), cables for opticalcommunication (ultrapure silica)
Translucency and chemical inertness	Heat- and corrosion-resistant materials, usually for Na lamps (Al_2O_3, MgO)
Nonlinearity	Switching devices for optical computing ($LiNbO_3$)
Infrared transparency	Infrared laser windows (CaF_2, SrF_2, NaCl)
Nuclear applications	
Fission	Nuclear fuel (UO_3, UC), fuel cladding (C, SiC), neutron moderators (C, BeO)
Fusion	Tritium breeder materials (zirconates and silicates of Li, Li_2O; fusion reactor lining (C, SiC, Si_3N_4, B_4C)
Chemical	
Catalysis	Filters (zeolites); purification of exhaust gases
Anticorrosion properties	Heat exchangers (SiC), chemical equipment in corrosive environment
Biocompatibility	Artificial joint prostheses (Al_2O_3)
Mechanical	
Hardness	Cutting tools (SiC whisker-reinforced Al_2O_3, Si_3N_4)
High-temperature strength retention	Stators and turbine blades, ceramic engines (Si_3N_4)
Wear resistance	Bearings (Si_3N_4)

of ceramics. The properties and applications of some important members are summarized in Table 3.55. For further reading the extensive treatment in [3.179] is recommended.

Nonoxide Ceramics

The *nonoxide ceramics* include essentially borides, carbides, nitrides, and silicides. A comprehensive overview of these materials is given in [3.1,177,178]. A few application examples will be given in the following. In recent years some effort has been made in the construction of ceramic automobile engine parts such as engine blocks, valves, cylinder liner, rotors for turbochargers, and so on. Ceramics under consideration for use in ceramic turbine engines include silicon nitride Si_3N_4, and silicon carbide SiC, which possess high thermal conductivity

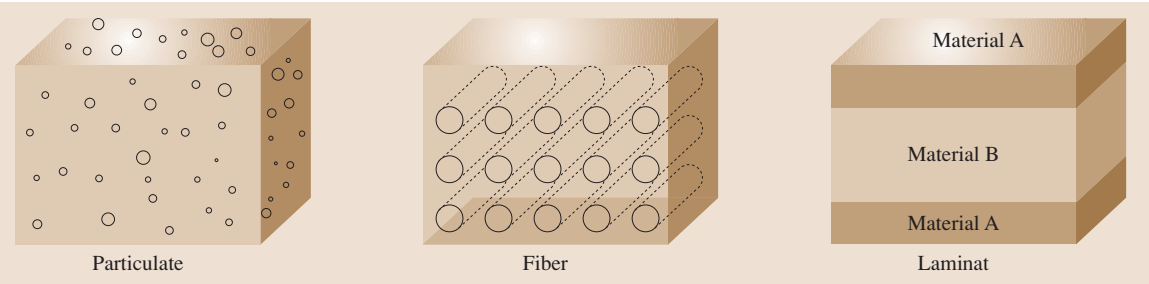


Fig. 3.171 Types of composites (schematic)

and thus excellent thermal-shock resistance. Boron carbide (B_4C), silicon carbide (SiC), and titanium diboride (TiB_2) are also being considered for armor systems to protect military personnel and vehicles from ballistic projectiles. The low density of ceramics makes them very attractive in this field. High-purity ceramics with simple crystal structures such as boron nitride (BN), silicon carbide (SiC), and aluminum nitride (AlN) may be used as substrate for integrated circuits (ICs), since they have better thermal conductivity and thermal expansion coefficients which are closer to the silicon IC chips than that of the presently used alumina. Further applications of nonoxide ceramics as well as advanced oxide ceramics are given in Table 3.56.

3.7.10 Composite Materials

Composite materials are formed when two materials which belong to different material classes are combined to attain properties which are not provided by the original materials. Possible combinations are:

- Metal–ceramics
- Metal–polymer
- Ceramic–polymer

The second phase could be introduced into the matrix material either in the form of homogeneously distributed particles, as fibers, or the materials form a laminate structure (Fig. 3.171).

In *dispersion-strengthened* alloys a small amount (usually $< 10\%$ volume fraction) of second-phase particles (metallic oxide or ceramic) are homogeneously distributed into the matrix material, commonly by mechanical alloying techniques [3.181]. These dispersoids are generally not coherent and, thus, effectively inhibit dislocation motion (Orowan circumvention at room temperature or by lowering dislocation line energy at higher temperatures [3.17]) when their distance is typically about 100 nm. The most important advantage over age-hardened alloys is the excellent elevated-temperature stability. Due to their insolubility within the metallic matrix no significant coarsening of the dispersoids is observed even after long-term exposure at temperatures close to the melting point of the matrix. Prominent examples are the oxide-dispersion-strengthened (ODS) nickel- and iron-based superalloys [3.182]. When the particles are large so that they do not significantly interact with moving dislocations, the rule of mixture can be applied for property determination, i. e., the properties can be directly determined by adding the percentage influence of each phase. Hence, the term *particulate-reinforced material* is more appropriate and the volume fraction of the reinforcement phase can exceed 50%. Applications are, e.g., cemented carbides as wear and cutting tools (Sect. 3.7.6), abrasives such as Al_2O_3, SiC, and BN, which are added to grinding and cutting wheels, and electrical contacts such as tungsten or oxide particle-reinforced silver.

The fibers in *fiber-reinforced* composites can be continuously (as in Fig. 3.171), orthogonal or randomly distributed. The properties of the fiber, i. e., strength, stiffness, etc., the aspect ratio l/d, where l is the fiber length and d is the diameter, and the volume fraction of

Table 3.57 Typical longitudinal and transverse tensile strengths for three unidirectionally fiber-reinforced composites. The fiber content in each is approximately 50 vol. % (after [3.180])

Material	Longitudinal tensile strength (MPa)	Transversale tensile strength (MPa)
Glass–polyester	700	20
Carbon (high modulus) –epoxy	1000	35
Kevlar–epoxy	1200	20

the fibers play a decisive role in the final performance of the reinforced composites. In the longitudinal direction (along the fiber axis) the strength is much higher than in the transverse direction (Table 3.57).

The matrix of fiber-reinforced materials should be tough enough to support the fibers and prevent cracks in broken fibers from propagating, and one has to be aware of chemical reactions when the matrix is a metallic material. If the fibers are exposed to high temperatures the coefficient of thermal expansion should not differ substantially from that of the matrix. Fiber composites may be used as fan blades in gas turbine engines and other aircraft and aerospace components, in lightweight automotive applications such as fiber-reinforced Al-matrix pistons, sporting goods (such as tennis rackets, golf club shafts, and fishing rods), and as corrosion-resistant components, to name some of the possible applications.

Laminar compositions could be very thin coatings such as thermal barrier coatings to protect Ni-based superalloys in high-temperature turbine applications (Sect. 3.7.5), thicker protective layers, or two-dimensional sheets or panels that have a preferred high-strength direction. The layers are stacked and joined by organic adhesives. Examples of laminar structures are adjacent wood sheets in plywood, capacitors composed of alternating layers of aluminum and mica, printed circuit boards, and insulation for motors, to mention a few.

References

3.1 W. Martienssen, H. Warlimont (eds.): *Springer Handbook of Condensed Matter and Materials Data* (Springer, Berlin, Heidelberg 2005)

3.2 R.J. Silbey, R.A. Alberty, M.G. Bawendi: *Physical Chemistry*, 4th edn. (Wiley, Hoboken 2005)

3.3 C. Kittel: *Introduction to Solid State Physics*, 8th edn. (Wiley, Chichester 2004)

3.4 R.W. Cahn, P. Haasen, E.J. Kramer: Materials Science and Technology. In: *Glasses and Amorphous Materials*, Vol. 9, ed. by J. Zarzycki (Wiley-VCH, Weinheim 2005)

3.5 A. Inoue, T. Zhang, T. Masumoto: Production of amorphous cylinder and sheet of $La_{55}Al_{25}Ni_{20}$ alloy by a metallic mold casting method, Mater. Trans. JIM **31**, 425 (1990)

3.6 A. Peker, W.L. Johnson: Example, Appl. Phys. Lett. **63**, 2342 (1993)

3.7 A. Leonhard, L.Q. Xing, M. Heilmaier, A. Gebert, J. Eckert, L. Schultz: Effect of crytalline precipitations on the mechanical behavior of bulk glass forming Zr-based alloys, Nanostructured Mater. **10**, 805 (1998)

3.8 A. Inoue: Stabilization of metallic supercooled liquid and bulk amorphous alloys, Acta Mater. **48**, 279 (2000)

3.9 R. Tilley: *Crystals and Crystal Structures* (Wiley, Chichester 2006)

3.10 R.W. Cahn, P. Haasen: *Physical Metallurgy*, Vol. 1–3, 4th edn. (North Holland, Amsterdam 1996)

3.11 D.A. Porter, K.E. Easterling: *Phase Transformations in Metals and Alloys*, 2nd edn. (Chapman Hall, London 1997)

3.12 D.B. Williams, C.B. Carter: *Transmission Electron Microscopy*, Vol. 1–4 (Plenum, New York 1996)

3.13 G.E. Dieter: *Mechanical Metallurgy, SI Metric Edition* (McGraw-Hill, London 1988)

3.14 J. Gurland: Stereology and Qualitative Metallography, ASTM. STP. **504**, 108 (1972)

3.15 E.O. Hall: The deformation and aging of mild steel: III discussion of results, Proc. Roy. Soc. B **64**, 747 (1951)

3.16 N.J. Petch: The cleavage strength of polycrystals, J. Iron Steel Inst. **174**, 25 (1953)

3.17 B. Reppich: Particle strengthening. In: *Materials Science and Technology*, Vol. 6 (Wiley-VCH, Weinheim 2005) pp. 312–357

3.18 H.E. Exner, M. Rettenmayr, C. Müller: Komplexe Grenzflächengeometrien bei Phasenumwandlungen, Prakt. Metallogr. **41**, 443–458 (2004)

3.19 V.K. Pecharsky, P. Zavalij: *Fundamentals of Powder Diffraction and Structural Characterization of Materials* (Kluwer Academic, Dordrecht 2003)

3.20 D.J. Dyson: *X-ray and Electron Diffraction Studies in Materials Science* (Maney, London 2004)

3.21 F.H. Chung, D.K. Smith: *Industrial Applications of X-ray Diffraction* (Dekker, New York 1999)

3.22 M. Howes, T. Inoue, G.E. Totten: *Handbook of Residual Stress and Deformation of Steel* (ASM Int, Materials Park 2002)

3.23 V. Randle, O. Engler: *Introduction to Texture Analysis* (Gordon Breach, Amsterdam 2000)

3.24 H.J. Bunge: *Texture Analysis in Materials Science, Mathematical Methods* (Butterworth-Heinemann, London 1982)

3.25 G.F. VanderVoort: *ASM Handbook: Metallography and Microstructures* (ASM Int., Materials Park 2004)

3.26 B. Bousfield: *Surface Preparation and Microscopy of Materials* (Wiley, Chichester 1992)

3.27 G. Petzow, V. Carle: *Metallographic Etching* (ASM Int., Materials Park 1999)

3.28 L. Reimer: *Scanning Electron Microscopy. Physics of Image Formation and Microanalysis* (Springer, Berlin 1998)

3.29 J.I. Goldstein, D.E. Newbury, P. Echlin, D.C. Joy, C. Fiori, E. Lifshin: *Scanning Electron Microscopy and X-ray Microanalysis* (Plenum, New York 1992)

3.30 W. Hauffe, R. Behrisch, K. Wittmaack: Sputtering by Particle Bombardment III, Topics in Appl. Phys. **64**, 111 (1991)

3.31 L.A. Giannuzzi, F.A. Stevie: *Introduction to Focused Ion Beams* (Springer, New York 2004)

3.32 B.W. Kempshall, S.M. Schwarz, B.I. Prenitzer, L.A. Giannuzzi, R.B. Irwin, F.A. Stevie: Ionchanneling effects on the focused ion beam milling of Cu, J. Vac. Sci. Technol. B **19**, 749–754 (2001)

3.33 L. Reimer: *Transmission Electron Microscopy. Physics of Image Formation and Microanalysis* (Springer, Berlin 2006)

3.34 C.C. Ahn, M.M. Disko, B. Fultz: *Transmission Electron Energy Loss Spectrometry in Materials Science and the EELS Atlas* (Wiley-VCH, Weinheim 2004)

3.35 J.P. Eberhart: *Structural and Chemical Analysis of Materials* (Wiley, Chichester 1991)

3.36 V.D. Scott, G. Love, S.J.B. Reed: *Quantitative Electron Probe Microanalysis* (Ellis Horwood, New York 1995)

3.37 A.J. Schwartz, M. Kumar, B.L. Adams: *Electron Backscatter Diffraction in Materials Science* (Kluwer Academic, New York 2000)

3.38 D.J. Dingley, K.Z. Baba-Kishi, V. Randle: *Atlas of Backscattering Kikuchi Diffraction Patterns* (IOP, Bristol 1995)

3.39 D. Katrakova, F. Mücklich: Specimen preparation for electron backscatter diffraction – Part I: Metals, Pract. Metallog. **38**, 547–565 (2001)

3.40 D. Katrakova, F. Mücklich: Specimen preparation for electron backscatter diffraction – Part II: Ceramics, Pract. Metallog. **39**, 644–662 (2002)

3.41 R.P. Goehner, J.R. Michael: Phase identification in a scanning electron microscope using backscattered electron kikuchi patterns, J. Res. Natl. Inst. Stand. Technol. **101**, 301 (1996)

3.42 R.A. Schwarzer, A. Huot: The study of microstructure on a mesoscale by ACOM, Cryst. Res. Technol. **35**, 851–862 (2000)

3.43 E.E. Underwood: *Quantitative Stereology* (Addison-Wesley, Reading 1970)

3.44 J. Ohser, F. Mücklich: *Statistical Analysis of Microstructures in Materials Science* (Wiley, Chichester 2000)

3.45 H.E. Exner: Quantitative description of microstructures by image analysis. In: *Materials Science and Technology*, Vol. 2B, ed. by R.W. Cahn, P. Haasen, E.J. Kramer (VCH, Weinheim 1994), pp 281–350

3.46 E.E. Underwood: *Metals Handbook*, Vol. 9, ed. by K Mills (ASM, Materials Park 1992) p. 123

3.47 ISO: *ISO 9042:1988: Steel – Manual Point Counting Method for Statistically Estimating the Volume Fraction of a Constituent with a Point Grid* (ISO, Geneva 1988)

3.48 ISO: *ISO 14250:2000 Steel – Metallographic Characterization of Duplex Grain Size and Distributions* (ISO, Geneva 2000)

3.49 ASTM: *Practice E1245-03: Standard Practice for Determining the Inclusion or Second-Phase Constituent Content of Metals by Automatic Image Analysis* (ASTM, Philadelphia 2003)

3.50 ISO: *ISO 945:1975: Cast Iron – Designation of Microstructure of Graphite* (ISO, Geneva 1975)

3.51 ISO: *ISO 2624:1990: Copper and Copper Alloys – Estimation of Average Grain Size* (ISO, Geneva 1990)

3.52 ASTM: *ASTM E112-96: Standard Test Methods for Determining Average Grain Size* (ASTM, Philadelphia 2004)

3.53 ISO: *ISO 643-2003: Steels – Micrographic Determination of the Apparent Grain Size* (ISO, Geneva 2003)

3.54 D.R. Askeland: *The Science and Engineering of Materials, S.I.*, 3rd edn. (Wiley-VCH, Weinheim 1996)

3.55 H.J. Frost, M.F. Ashby: *Deformation Mechanism Maps* (Oxford Univ. Press, Oxford 1982)

3.56 O.D. Sherby, J. Wadsworth: Superplasticity – Recent advances and future directions, Prog. Mater. Sci. **33**, 169 (1989)

3.57 S. Suresh: *Fatigue of Materials* (Cambridge Univ. Press, Cambridge 1998)

3.58 Annual Book of ASTM Standards, Vol. 03.01 (ASTM, Philadelphia 1998)

3.59 J. Buschow: *Concise Encyclopedia of Magnetic and Superconducting Materials*, 2nd edn. (Elsevier, Amsterdam 2006)

3.60 S. Ness, C.N. Sherlock, P.O. Moore, P. McIntire: *NDT Handbook*, Overview, Vol. 10, 2nd edn. (ASNT, Columbus 1996)

3.61 H. Blumenauer: *Werkstoffprüfung* (Deutscher Verlag Grundstoffindustrie, Leipzig Stuttgart 1994), in German

3.62 W. Grellmann, S. Seidler: *Kunststoffprüfung* (Carl-Hanser, München Wien 2005), in German

3.63 K. Nitsche: *Schichtmeßtechnik* (Vogel, Würzburg 1997), in German

3.64 S. Steeb: *Zerstörungsfreie Werkstoff- und Werkstückprüfung* (Expert, Ehningen 2004), in German

3.65 J. Krautkrämer, H. Krautkrämer: *Ultrasonic Testing of Materials* (Springer, Berlin, Heidelberg 1990)

3.66 A.S. Birks, R.E. Green Jr., P. McIntire: *Ultrasonic Testing*, NDT Handbook, Vol. 7, 2nd edn. (ASNT, Columbus 1991)

3.67 U. Netzelmann, H. Reiter, Y. Shi, J. Wang, M. Maisl: Ceramic automotive valves – Chances and limitations of nondestructive testing, e-J. NDT **2**, 7 (1997), http://www.ndt.net

3.68 I. Hertlin, T. Herkel: *Riss- und Gefügeprüfung mit akustischer Resonanzanalyse im Schall- und Ultraschallbereich für Kfz-Sicherheitsteile* (Annu. Conf. DGZfP, Mainz 2003), V18

3.69 R.K. Miller, E.V.K. Hill: *Acoustic Emission Testing*, NDT Handbook, Vol. 6, 3rd edn. (ASNT, Columbus 2005)

3.70 S. Roderick, P.O. Moore, P. McIntire: *Special Nondestructive Methods*, NDT Handbook, Vol. 9, 2nd edn. (ASNT, Columbus 1995)

3.71 B.G. Livschitz: *Physikalische Eigenschaften der Metalle und Legierungen* (Deutscher Verlag Grundstoffindustrie, Leipzig 1989), in German

3.72 J.T. Schmidt, K. Skeie, P. McIntire: *Magnetic Particle Testing*, NDT Handbook, Vol. 6, 2nd edn. (ASNT, Columbus 1989)

3.73 H. Heptner, H. Stroppe: *Magnetische und magnetinduktive Werkstoffprüfung* (Deutscher Verlag Grundstoffindustrie, Leipzig 1972), in German

3.74 W. Morgner, F. Michel: *Some New Results in Nondestructive Case Depth Measurement* (9th European Conference on NDT, Berlin 2006), Paper 118

3.75 W.D. Feist, G. Mook, J.H. Hinken, J. Simonin, H. Wrobel: Electromagnetic detection and characterization of tungsten carbide inclusions in non-ferromagnetic alloys, Adv. Eng. Mat. **7**(9), 841–846 (2005)

3.76 W. Willmann, G. Wollmann: Der Barkhausen-Effekt und seine technische Nutzung, Exp. Techn. Phys. **31**, 533–543 (1983)

3.77 G. Dobmann: Nondestructive Testing of Laser Processing of Material, Workshop on Laser Techniques (2003)

3.78 I. Altpeter, J. Bender, J. Hoffmann, D. Rouget: *Barkhausen-Effect and Eddy-Current Testing for the Characterization of the Microstructure and Residual Stress States with Local Resolution, In: EURO MAT '97: Characterization and Production/Design*, Vol. 4 (Society for Materials Science, Zwijndrecht 1997) pp. 123–128

3.79 S. Udpa, P.O. Moore: *Electromagnetic Testing*, NDT Handbook, Vol. 5, 3rd edn. (ASNT, Columbus 2004)

3.80 R. Zoughi: *Microwave Nondestructive Testing and Evaluation – A Graduate Textbook* (Kluwer Academic, Dordrecht 2000)

3.81 A.J. Bahr: *Microwave Nondestructive Testing Methods* (Gordon Breach, New York 1982)

3.82 G. Busse: Zerstörungsfreie Kunststoffprüfung. In: *Kunststoffprüfung*, ed. by W. Grellmann, S. Seidler (Carl-Hanser, München, Wien 2005)

3.83 X.V.P. Maldague, P.O. Moore: *Infrared and Thermal Testing*, NDT Handbook, Vol. 3, 3rd edn. (ASNT, Columbus 2001)

3.84 S.R. Lampman, T.B. Zorc, H.J. Frissell, G.M. Crankovic, A.W. Ronke: *Nondestructive Evaluation and Quality Control*, ASM Handbook 17 (ASM International, Materials Park 1989)

3.85 N. Tracy, P.O. Moore: *Liquid Pentrant Testing*, NDT Handbook, Vol. 2, 3rd edn. (ASNT, Columbus OH 1999)

3.86 W.J. Bisle, D. Scherling, M.K. Kalms, W. Osten: Improved shearography for use on optical non cooperating surfaces under daylight conditions, AIP Conf. Proc. **557**(1), 1928–1935 (2001)

3.87 R.H. Bossi, F.A. Iddings, G.C. Wheeler, P.O. Moore: *Radiographic Testing*, NDT Handbook, Vol. 4, 3rd edn. (ASNT, Columbus 2002)

3.88 R. Glocker: *Materialprüfung mit Röntgenstrahlen* (Springer, Berlin, Heidelberg 1985), in German

3.89 G. Mook, J. Pohl, F. Michel: Non-destructive characterization of smart CFRP structures, Smart Mater. Struct. **12**, 997–1004 (2003)

3.90 J. Pohl, S. Herold, G. Mook, F. Michel: Damage detection in smart CFRP composites using impedance spectroscopy, Smart Mater. Struct. B **10**, 834–842 (2001)

3.91 H. Speckmann, R. Henrich: *Structural Health Monitoring (SHM) – Overview on Airbus Activities* (16th World Conf. NDT, Montreal 2004), paper 536

3.92 A.K. Mukherjee, J.E. Bird, J.E. Dorn: Experimental correlations for high-temperature creep, Trans. ASM **62**, 155 (1969)

3.93 W.J. Staszewski, C. Boller, G.R. Tomlinson: *Health Monitoring of Aerospace Structures: Smart Sensor Technologies and Signal Processing* (Wiley, New York 2003)

3.94 F.-K. Chang (Ed.): *Structural Health Monitoring. The Demands and Challenges* (CRC Press, Boca Raton 2002)

3.95 P.R. Roberge: *Corrosion Basics – An Introduction* (NACE International, Houston 2006)

3.96 K.A. van Oeteren: *Korrosionsschutz durch Beschichtungsstoffe* (Hanser, München 1980), in German

3.97 P. Maaß, P. Peißker: *Handbuch Feuerverzinken* (Deutscher Verlag Grundstoffindustrie, Stuttgart 1970), in German

3.98 U.R. Evans: Some recent work on the corrosion of metals, Metal Ind. **29**, 481 (1926)

3.99 H. Baum: *Untersuchungen zum Mechanismus der Deckschichtbildung beim atmosphärischen Rosten korrosionsträger Stähle* Dissertation, Bergakademie Freiberg (1973)

3.100 Institut für Korrosionsschutz Dresden: *Vorlesungen über Korrosion und Korrosionsschutz* (TAW, Wuppertal 1996)

3.101 J. Göllner: Elektrochemisches Rauschen unter Korrosionsbedingungen, Habilitation, Otto-von-Guericke-Universität Magdeburg (2002)

3.102 T. Shibata: Stochastic approach to the effect of alloying elements on the pitting resistance of ferritic stainless steels, Trans. ISIJ, **23**, 785–788 (1983)

3.103 H.H. Uhlig: *Corrosion and Corrosion Control* (Wiley, New York 1971)

3.104 K. Mörbe, W. Morenz, H.-W. Pohlmann, H. Werner: *Korrosionsschutz wasserführender Anlagen* (Springer, Wien 1998)

3.105 K.H. Tostmann: *Korrosion* (Verlag Chemie, Weinheim 2001)

3.106 E. Wendler-Kalsch, H. Gräfen: *Korrosionsschadenskunde* (Springer, Berlin 1998), in German

3.107 K. Schilling: *Selektive Korrosion hochlegierter Stähle*, Dissertation, Otto-von-Guericke-Universität Magdeburg (2005)

3.108 H. Kaesche: *Corrosion of Metals* (Springer, Berlin 2003)

3.109 C. Wagner, W. Traud: Über die Deutung von Korrosionsvorgängen durch Überlagerung von elektrochemischen Teilvorgängen und über die Poten-

tialbildung an Mischelektroden, Z. Elektroch. **44**, 391–454 (1938), in German
3.110 E. Hornbogen, H. Warlimont: *Metallkunde* (Springer, Berlin 2001), in German
3.111 R.W. Cahnn, P. Haasen, E.J. Kramer, M. Schütze: *Corrosion and Environmental Degradation*, Materials Science and Technology (Wiley-VCH, Weinheim 2000)
3.112 W. Schatt, H. Worch: *Werkstoffwissenschaft* (Deutscher Verlag Grundstoffindustrie, Stuttgart 1996)
3.113 R.B. Ross: *Metallic Materials Specification Handbook*, 4th edn. (Chapman Hall, London 1992)
3.114 A. Nayar: *The Metals Databook* (McGraw-Hill, New York 1997)
3.115 M.F. Ashby, D.R.H. Jones: *Engineering Materials 2: An Introduction to Microstructures, Processing and Design* (Butterworth-Heinemann, Burlington 1998)
3.116 A.M. Howatson, P.G. Lund, J.D. Todd: *Engineering Tables and Data*, 2nd edn. (Chapman Hall, London 1991)
3.117 D.K. Roylance: Mechanics of Materials, Massachusetts Institute of Technology Department of Materials Science and Engineering, Cambridge (MIT-DMSE), Material Properties (http://web.mit.edu/course/3/3.11/www/modules/props.pdf)
3.118 R.W.K. Honeycombe, H.K.D.H. Bhadeshia: *Steels – Microstructure and Properties*, 2nd edn. (Edward Arnold, London, New York, Sydney, Auckland 1995)
3.119 G. Krauss: *Steel – Heat Treatment and Processing Principles* (ASM Int., Materials Park 1989)
3.120 W.C. Leslie: *The Physical Metallurgy of Steels* (McGraw-Hill, New York 1981)
3.121 A.K. Sinha: *Ferrous Physical Metallurgy* (Butterworths, London 1989)
3.122 D.T. Llewellyn, R.C. Hudd: *Steels, Metallurgy and Applications* (Butterworth Heinemann, Oxford 1998)
3.123 Online Source: Key to Steel: Steel Database on http://www.key-to-steel.com/
3.124 H.K.D.H. Bhadeshia: *Bainite in Steels, Transformation, Microstructure and Properties* (IOM, London 2001)
3.125 J.R. Davis: *Carbon and Alloy Steels*, ASM Speciality Handbook (ASM, Metals Park 1996)
3.126 E.C. Bain, H.W. Paxton: *Alloying Elements in Steel* (ASM, Metals Park 1966)
3.127 P.M. Unterweiser: *Worldwide Guide to Equivalent Irons and Steels* (ASM, Materials Park 1996)
3.128 J.E. Bringes: *Handbook of Comparative World Steel Standards* (ASTM, West Conshohocken 2001)
3.129 J.R. Davis: *Stainless Steels*, ASM Speciality Handbook (ASM, Metals Park 1994)
3.130 J.R. Davis: *Tool Materials*, ASM Specialty Handbook (ASM, Metals Park 1995)
3.131 H.E. McGannon: *The Making, Shaping and Treatment of Steel* (United States Steel Corporation, Pittsburgh 1971)
3.132 J.R. Davis: *Cast Irons*, ASM Specialty Handbook (ASM, Metals Park 1996)
3.133 W.G. Moffatt, G.W. Pearsall, J. Wulff: *The Structure and Properties of Materials*, Structure, Vol. 1 (Wiley, New York 1964), p. 195
3.134 Specialty Castings Inc. http://www.specialtycastings.com/ductile_iron.html
3.135 G.E. Totten, D.S. MacKenzie: *Physical Metallurgy and Processes*, Handbook of Aluminum, Vol. 1 (Dekker, New York 2003)
3.136 C. Kammer: *Fundamentals and Metarials*, Aluminium Handbook 1 (Aluminium Verlag, Düsseldorf 2002), in German
3.137 J.R. Davis (Ed.): *Aluminum and Aluminum Alloys*, ASM Specialty Handbook (ASM, Metals Park 1993)
3.138 The University of British Columbia, Department of Materials Engineering, Mmat 380: online course material, Heat treatable aluminium alloys, http://www.mmat.ubc.ca/courses/mmat380/default.htm
3.139 A. Dehler, S. Knirsch, V. Srivastava, H. Saage, M. Heilmaier: Assessment of creep behaviour of the die-cast cylinder-head alloy AlSi6Cu4-T6, Int. J. Met. Res. **97**, 12 (2006)
3.140 H. Baker, B. David, K.W. Craig: *Metals Hanbook*, Vol. 2 (ASM, Metal Parks 1979)
3.141 M.M. Avdesian, H. Baker: *Magnesium and Magnesium Alloys*, ASM Specialty Handbook (ASM, Metals Park 1999)
3.142 I.J. Polmear: *Light Alloys, Metallurgy of the Light Metals* (Wiley, New York 1995)
3.143 G. Neite: Structure and properties of nonferrous alloys. In: *Materials Science and Technologie*, Vol. 8, ed. by K.H. Matucha (Verlag Chemie, Weinheim 1996)
3.144 The University of British Columbia, Department of Materials Engineering – mmat 380: online course material, Titanium alloys, http://www.mmat.ubc.ca/courses/mmat380/default.htm
3.145 R. Boyer, G. Welsch, E.W. Collings: *Materials Properties Hanbook: Titanium Alloys* (ASM, Materials Park 1994)
3.146 K.H. Matchuta: Structure and properties of nonferrous alloys. In: *Matreials Science and Technology*, Vol. 8, ed. by R.W. Cahn, P. Haasen, E.J. Kramer (VCH, Weinheim 1996)
3.147 W.F. Hosford: *Physical Metallurgy* (Taylor Francis, New York 2005)
3.148 S.C. Huang, J.C. Chessnut: *Intermetallic Compounds-Principles and Practice, Vol. 2*, Vol. 2, ed. by J.H. Westbrook, R.L. Fleischer (Wiley, Chinchester 1994) p. 73
3.149 Forschungszentrum Jülich GmbH: *Titan-Aluminid-Legierungen – eine Werkstoffgruppe mit Zukunft* (Grafische Betriebe, Forschungszentrum Jülich GmbH, Jülich 2003), in German
3.150 K. Otsuka, C.M. Wayman: *Shape Memory Materials* (Cambridge Univ. Press, Cambridge 1998)

3.151 J.R. Davies: *Heat-Resistant Materials*, ASM Specialty Handbook (ASM Int., Metals Park 1997)
3.152 G. Joseph, K.J.A. Kundig: *Copper, Its Trade, Manufacture, Use, and Environment Status* (ASM Int., Materials Park 1998)
3.153 J.R. Davis: *Copper and Copper Alloys*, ASM Specialty Handbook (ASM, Metals Park 2001)
3.154 H. Lipowsky, E. Arpaci: *Copper in the Automotive Industry* (Wiley-VCH, Weinheim 2006)
3.155 J. Brandrup, E.H. Immergut, E.A. Grulke: *Polymer Handbook* (Wiley, New York 2004)
3.156 H.-G. Elias: *An Introduction to Polymer Science* (Wiley-VCH, Weinheim 1999)
3.157 I. Mita, R.F.T. Stepto, U.W. Suter: Basic classification and definitions of polymerization reactions, Pure Appl. Chem. **66**, 2483–2486 (1994)
3.158 K. Matyjaszewski, T.P. Davis: *Handbook of Radical Polymerization* (Wiley, New York 2002)
3.159 G.W. Ehrenstein, R.P. Theriault: *Polymeric Materials: Structure, Properties, Applications* (Hanser Gardner, Munich 2000)
3.160 G.H. Michler, F.J. Baltá-Calleja: *Mechanical Properties of Polymers Based on Nano-Structure and Morphology* (CRC, Boca Raton 2005)
3.161 A.E. Woodward: *Atlas of Polymer Morphology* (Hanser Gardner, Munich 1988)
3.162 E.A. Campo: *The Complete Part Design Handbook for Injection Moulding of Thermoplastics* (Hanser, Munich 2006)
3.163 D.V. Rosato, A.V. Rosato, D.P. DiMattia: *Blow Moulding Handbook* (Hanser Gardner, Munich 2003)
3.164 L.C.E. Struik: *Internal Stresses, Dimensional Instabilities and Molecular Orientations in Plastics* (Wiley, New York 1990)
3.165 ISO: *ISO 1135 parts 1-7:1997: Plastics – Differential Scanning Calorimetry (DSC) – Part 1: General Principles* (ISO, Geneva 1997)
3.166 T.A. Osswald, G. Menges: *Materials Science of Polymers for Engineers* (Hanser, Munich 1995)
3.167 P.C. Powell: *Engineering with Polymers* (CRC, Boca Raton 1998)
3.168 I.M. Ward, D.W. Hadley: *An Introduction to the Mechanical Properties of Solid Polymers* (Wiley, Chichester 1993)
3.169 H. Czidios, T. Saito, L. Smith (Eds.): *Springer Handbook of Materials Measurement Methods* (Springer, Berlin, Heidelberg 2006), Chap. 7
3.170 I.M. Ward: *Structure and Properties of Oriented Polymers* (Chapman Hall, London 1997)
3.171 ISO: *ISO 6721-1:2001 Plastics – Determination of Dynamic Mechanical Properties – Part 1: General Principles; ISO 6721-2: 1994 Plastics – Determination of Dynamic Mechanical Properties – Part 2: Torsion-Pendulum Method* (ISO, Geneva 2001)
3.172 E.A. Grulke: Solubility parameter values. In: *Polymer Handbook 3rd. edn*, ed. by J. Brandrup, E.H. Immergut (Wiley, New York 1989), VII/519–557
3.173 G.W. Ehrenstein: *Faserverbund-Kunststoffe, Werkstoffe – Verarbeitung – Eigenschaften* (Hanser, Munich 2006)
3.174 L.H. Sperling: *Polymeric Multicomponent Materials* (Wiley, New York 1997)
3.175 C.M. Hansen: *Solubility Parameters: A User's Handbook* (CRC, Boca Raton 1999)
3.176 W.D. Callister Jr.: *Fundamentals of Materials Science and Engineering* (Wiley, New York 2001)
3.177 R. Freer: *The Physics and Chemistry of Carbides, Nitrides and Borides* (Kluwer, Boston 1989)
3.178 M.V. Swain: *Structure and Properties of Ceramics*, Materials Science and Technology, Vol. 11 (Verlag Chemie, Weinheim 1994)
3.179 G.V. Samson: *The Oxides Handbook* (Plenum, New York 1974)
3.180 D. Hull, T.W. Clyne: *An Introduction to Composite Materials*, 2nd edn. (Cambridge Univ. Press, Cambridge 1996)
3.181 J.S. Benjamin: Dispersion strengthened superalloys by mechanical alloying, Metall. Trans. **1**, 2943 (1970)
3.182 Y. Estrin, S. Arndt, M. Heilmaier, Y. Brechet: Deformation beahviour of particle strengthened alloys: A Voronoi mesh approach, Acta Mater. **47**, 595 (1999)

4. Thermodynamics

Frank Dammel, Jay M. Ochterbeck, Peter Stephan

This chapter presents the basic definitions, laws and relationships concerning the thermodynamic states of substances and the thermodynamic processes. It closes with a section describing the heat transfer mechanisms.

4.1 Scope of Thermodynamics. Definitions

Thermodynamics is a subsection of physics that deals with energy and its relationship with properties of matter. It is concerned with the different forms of energy and their transformation between one another. It provides the general laws that are the basis for energy conversion, transfer, and storage.

4.1.1 Systems, System Boundaries, Surroundings

A thermodynamic system, or briefly a system, is a quantity of matter or a region in space chosen for a thermodynamic investigation. Some examples of systems are an amount of gas, a liquid and its vapor, a mixture of several liquids, a crystal or a power plant. The system is separated from the surroundings, the so-called environment, by a boundary (real or imaginary). The boundary is allowed to move during the process under investigation, e.g., during the expansion of a gas, and matter and energy may cross the boundary. Energy can cross a boundary with matter and in the form of heat transfer or work (Sect. 4.3.2). The system with its boundary serves as a region with a barrier in which computations of energy conversion processes take place. Using an energy balance relationship (the first law of thermodynamics) applied to a system, energies that cross the system boundary (in or out), the changes in stored energy, and the properties of the system are linked. A system is called closed when mass is not allowed to cross the boundary, and open when mass crosses the system boundary. While the mass of a closed system always remains constant, the mass inside an open system may also remain constant when the total mass flow in and the total mass flow out are equal. Changes of the mass stored in an open system will occur when the mass flow into the system over a certain time span is different from the mass flow out. Examples of closed systems are solid bodies, mass elements in mechanics, and a sealed container. Examples of open systems are turbines, turbojet engines, or a fluid (gases or liquids) flowing in channel. A system is called adiabatic when it is completely thermally isolated from its surrounding and no heat transfer can cross the boundary. A system that is secluded from all influences of its environment is called isolated. For an isolated system neither energy in the form of heat transfer or work nor matter are exchanged with the environment.

The distinction between a closed and an open system corresponds to the distinction between a Lagrangian and an Eulerian reference system in fluid mechanics. In the Lagrangian reference system, which corresponds to the closed system, the fluid motion is examined by dividing the flow into small elements of constant mass and deriving the corresponding equations of motion. In the Eulerian reference system, which corresponds to the open system, a fixed volume element in space is selected and the fluid flow through the volumetric element is examined. Both descriptions are equivalent, and it is often only a question of convenience whether one chooses a closed or an open system.

4.1.2 Description of States, Properties, and Thermodynamic Processes

A system is characterized by physical properties, which can be given at any instant, for example, pressure, temperature, density, electrical conductivity, and refraction index. The state of a system is determined by the values of these properties. The transition of a system from one equilibrium state to another is called a change of state.

Example 4.1: A balloon is filled with gas. The gas may then be the thermodynamic system. Measurements show that the mass of the gas is determined by volume, pressure, and temperature. The properties of the system are thus volume, pressure, and temperature, and the state of the system (the gas) is characterized through a fixed set of volume, pressure, and temperature. The transition to another fixed set, e.g., when a certain amount of gas effuses, is called a change of state.

The mathematical relationship between properties is called an equation of state.

Example 4.2: The volume of the gas in the balloon proves to be a function of pressure and temperature. The mathematical relationship between these properties is such an equation of state.

Properties are subdivided into three classes: *intensive properties* are independent of the size of a system and thus keep their values after a division of the system into subsystems.

Example 4.3: If a space filled with a gas of uniform temperature is subdivided into smaller spaces, the temperature remains the same in each subdivided space. Thus, temperature is an intensive property. Pressure would be another example of an intensive property.

Properties that are proportional to the mass of the system (i. e., the total is equal to the sum of the parts) are called *extensive properties*.

Example 4.4: The volume, the energy or the mass.

An extensive property X divided by the mass m of the system yields the *specific property* $x = X/m$.

Example 4.5: Take the extensive property volume of a given gas. The associate specific property is the specific volume $v = V/m$, where m is the mass of the gas. The SI unit for specific volume is m^3/kg. Specific properties all fall into the category of intensive properties.

Changes of state are caused by interactions of the system with the environment, for example, when energy is transferred to or from the system across the system boundary. In order to describe a change of state it is sufficient to specify the time history of the properties. The description of a process requires additional specifications of the extent and type of the interactions with the environment. Consequently, a process is a change of state caused by certain external influences. The term *process* is more comprehensive than the term *change of state*; for example, the same change between two states can be induced by different processes.

4.2 Temperatures. Equilibria

4.2.1 Thermal Equilibrium

We often talk about *hot* or *cold* bodies without quantifying such states exactly by a property. When a closed hot system A is exposed to a closed cold system B, energy is transported as heat transfer through the contact area. Thereby, the properties of both systems change until after a sufficient period of time new fixed values are reached and the energy transport stops. The two systems are in *thermal equilibrium* in this final state. The speed with which this equilibrium state is approached depends on the type of contact between the two systems and on the thermal properties. If, for example, the two systems are separated only by a thin metal wall, the equilibrium is reached faster than in the case of a thick polystyrene wall. A separating wall, which inhibits mass transfer and also mechanical, magnetic or electric interactions, but permits the transport of heat, is called diatherm. A diatherm wall is *thermally* conductive. A completely thermally insulated wall such that no thermal interactions occur with the surroundings is called adiabatic.

4.2.2 Zeroth Law and Empirical Temperature

In the case of thermal equilibrium between systems A and C and thermal equilibrium between systems B and C experience shows that the systems A and B must also be in thermal equilibrium. This empirical statement is called the *zeroth law of thermodynamics*. It reads: if two systems are both in thermal equilibrium with a third system, they are also in thermal equilibrium with each other. In order to find out if two systems A and B are in thermal equilibrium, they are exposed successively to a system C. The mass of system C may be small compared to those of systems A and B. If so, changes in state of systems A and B are negligible during equilibrium adjustment. When C is exposed to A, certain properties of C will change, for example, its electrical resistance. These properties then remain unchanged during the following exposure of C to B, if A and B were originally in thermal equilibrium. Using C in this way it is possible to verify if A and B are in thermal equilibrium. It is possible to assign any fixed values to the properties of C after equilibrium adjustment. These values are called *empirical temperatures*, where the measurement instrument is a thermometer.

4.2.3 Temperature Scales

A gas thermometer (Fig. 4.1), which measures the pressure p of a constant gas volume V, is used for the construction and definition of empirical temperature scales. The gas thermometer is brought into contact with systems of a constant state, e.g., a mixture of ice and water at a fixed pressure. After a sufficient period of time, the gas thermometer will be in thermal equilibrium with the system with which it is in contact. The gas volume is kept constant by changing the height Δz of the mercury column. The pressure exerted by the mercury column and environment is measured and the product pV is computed. The extrapolation of measurements at different, sufficiently low pressures leads to a threshold value A of the product pV for the pressure approaching zero. This value A, which is determined from the measurements, is assigned to an empirical temperature via the linear relationship

$$T = \text{const.}\, A\ . \tag{4.1}$$

After fixing the value *const* it is only necessary to determine the value of A from the measurements in order to compute the empirical temperature with (4.1). The specification of the empirical temperature scale requires

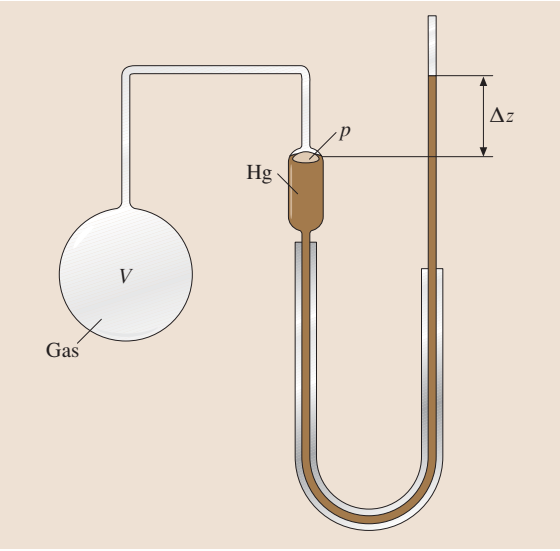


Fig. 4.1 Gas thermometer with gas volume in piston and mercury column

a fixed point. The 10th General Conference of Weights and Measures, held in Paris in 1954, assigned the triple point of water to a temperature $T_{tr} = 273.16$ Kelvin (designated by K). At the triple point of water vapor, liquid water, and ice coexist in equilibrium at a pressure of 611.657 ± 0.010 Pa. The temperature scale introduced in this way is named the *Kelvin scale*, and it is identical to the thermodynamic temperature scale. It holds that

$$T = T_{tr} A / A_{tr} , \tag{4.2}$$

if A_{tr} is the value of A measured with a gas thermometer at the triple point of water.

On the *Celsius scale*, where the unit of temperature t is designated by °C, the ice and steam points are assigned the values of $t_0 = 0\,°C$ and $t_1 = 100\,°C$, respectively, at a pressure of 0.101325 MPa. This corresponds quite accurately to absolute temperatures of $T_0 = 273.15\,K$ and $T_1 = 373.15\,K$. The temperature $T_{tr} = 273.16\,K$ at the triple point of water is roughly 0.01 K higher than the temperature at the ice point. The conversion of temperatures is carried out according to the equation

$$T = t + 273.15 , \tag{4.3}$$

where t is in °C and T is in K.

Additionally, the Fahrenheit scale is common in some countries, particularly the USA. The corresponding values on this scale are 32°F at the ice point and 212°F at the steam point of water (pressure in each case 0.101325 MPa). Conversion of a temperature t_F given in °F to a Celsius temperature t in °C is given by

$$t = \frac{5}{9}(t_F - 32) . \tag{4.4}$$

The degree increments of the Rankine scale (°R) are the same as Fahrenheit degrees, however, the reference 0 is set at absolute zero. It holds that

$$T_R = \frac{9}{5} T , \tag{4.5}$$

where T_R is in °R and T is in K. The ice point of water is thus given as 491.67 °R.

The International Practical Temperature Scale

Since it is difficult and time consuming to measure temperatures precisely with a gas thermometer, the international practical temperature scale was introduced by law. It is arranged by the International Committee for Weights and Measures so that its temperature approaches as close as possible the thermodynamic temperature of certain substances. The international practical temperature scale is fixed by the freezing and boiling points of these substances, which were determined as precisely as possible with a gas thermometer by the scientific national institutes of the different countries. Resistance thermometers, thermocouples, and radiation measuring devices are used to interpolate between the fixed points, whereas certain instructions are given for the relationships between the actually measured quantities and the temperature. The basic regulations of the international temperature scale are the same in all countries. They read:

1. In the international temperature scale of 1948 the symbol of the temperature is t and its unit is "°C" or "°C (Int. 1948)".
2. On the one hand the scale is based on a number of always reproducible equilibrium temperatures (fixed points), which are assigned to certain numerical values, and on the other hand on accurately defined formulas, which establish relationships between the temperature and the indications of the measuring instruments calibrated at the fixed points.
3. The fixed points and the assigned numerical values are summarized in tables (Table 4.1). With the exception of the triple points the assigned temperatures correspond to equilibrium states at the pressure 0.101325 MPa, which is the standard atmospheric pressure at sea level.

Table 4.1 Fixed points of the international temperature scale of 1990 (IPTS-90)

Equilibrium state	Assigned values of the international practical temperature scale	
	T_{90} (K)	t_{90} (°C)
Vapor pressure of helium	3 to 5	−270.15 to −268.15
Triple point of equilibrium hydrogen	13.8033	−259.3467
Vapor pressure of equilibrium hydrogen	≈ 17	≈ −256.15
	≈ 20.3	≈ −252.85
Triple point of neon	24.5561	−248.5939
Triple point of oxygen	54.3584	−218.7916
Triple point of argon	83.8058	−189.3442
Triple point of mercury	234.3156	−38.8344
Triple point of water	273.16	0.01
Melting point of gallium	302.9146	29.7646
Solidification point of indium	429.7485	156.5985
Solidification point of tin	505.078	231.928
Solidification point of zinc	692.677	419.527
Solidification point of aluminium	933.473	660.323
Solidification point of silver	1234.93	961.78
Solidification point of gold	1337.33	1064.18
Solidification point of copper	1357.77	1084.62

All substances beside helium may have their natural isotope composition. Hydrogen consists of ortho- and parahydrogen at equilibrium composition.

Table 4.2 Some thermometric fixed points: E solidification point, Sd boiling point at pressure 101.325 kPa, Tr triple point (after [4.1])

		°C
Normal hydrogen	Tr	−259.198
Normal hydrogen	Sd	−252.762
Nitrogen	Sd	−195.798
Carbon dioxide	Tr	−56.559
Bromine benzene	Tr	−30.726
Water (saturated with air)	E	0
Benzoic acid	Tr	122.34
Indium	Tr	156.593
Bismuth	E	271.346
Cadmium	E	320.995
Lead	E	327.387
Mercury	Sd	356.619
Sulfur	Sd	444.613
Antimony	E	630.63
Palladium	E	1555
Platinium	E	1768
Rhodium	E	1962
Iridium	E	2446
Tungsten	E	3418

4. Formulas, which also are established by international agreements, are used for interpolation between fixed points. Thus, the indications of the standard instruments with which the temperatures have to be measured, are assigned to the numerical values of the international practical temperature.

In order to simplify temperature measurements other additional thermometric fixed points for substances, which can be easily produced in sufficiently pure form, were associated as accurately as possible to the lawful temperature scale. The most important ones are summarized in Table 4.2. The platinum resistance thermometer is used as the normal instrument between the triple point of equilibrium hydrogen at 13.8033 K (−259.3467 °C) and the melting point of silver at 1234.93 K (961.78 °C). Between the melting point of silver and the melting point of gold at 1337.33 K (1064.18 °C) a platinum–rhodium (10% rhodium)/platinum thermocouple is used as normal instrument. Above the melting point of gold, Planck's radiation law defines the international practical temperature

$$\frac{J_t}{J_{\mathrm{Au}}} = \frac{\exp\left(\frac{c_2}{\lambda(t_{\mathrm{Au}}+T_0)}\right)-1}{\exp\left(\frac{c_2}{\lambda(t+T_0)}\right)-1}\,, \qquad (4.6)$$

where J_t and J_{Au} are the radiation energies emitted by a black body at temperature t and at the gold point t_{Au}, respectively, at a wavelength of λ per unit area, time, and wavelength interval. The value of the constant c_2 is specified as 0.014388 Km (Kelvin meter), $T_0 = 273.15$ K is the numerical value of the melting temperature of ice, and λ is the numerical value in m of a wavelength in the visible spectrum. For practical temperature measurement [4.2, 3]

4.3 First Law of Thermodynamics

4.3.1 General Formulation

The first law is an empirical statement, which is valid because all conclusions drawn from it are consistent with experience. Generally, it states that energy can be neither destroyed nor created, thus energy is a conserved property. This means that the energy E of a system can be changed only by energy exchange into or out of the system. It is generally agreed that energy transferred to a system is positive and energy transferred from a system is negative. A fundamental formulation of the first law reads: every system possesses an extensive property energy, which is constant in an isolated system.

4.3.2 The Different Forms of Energy and Energy Transfer

In order to set up the first law mathematically it is necessary to distinguish and define the different forms of energy transfer.

Work

In thermodynamics the basic definition of the term *work* is adopted from mechanics: the work done on a system is equal to the product of the force acting on the system and the displacement from the point of application. The work done by a force F along the distance z between points 1 and 2 is given by

$$W_{12} = \int_1^2 \boldsymbol{F}\,\mathrm{d}z\,. \tag{4.7}$$

The mechanical work W_{m12} is the result of forces which accelerate a closed system of mass m from velocity w_1 to w_2 and raise it from level z_1 to level z_2 against gravity g. This associates a change in kinetic energy $mw^2/2$ and in potential energy mgz of the system

$$W_{m12} = m\left(\frac{w_2^2}{2} - \frac{w_1^2}{2}\right) + mg(z_2 - z_1)\,. \tag{4.8}$$

Equation (4.8) is known as the energy theorem of mechanics.

Moving boundary work, or simply boundary work, is the work that has to be done to change the volume of a system. In a system of volume V, which possesses the variable pressure p, a differential element $\mathrm{d}A$ of the boundary surface thereby moves the distance $\mathrm{d}z$. The work done is

$$\mathrm{d}W_v = -p\int_A \mathrm{d}\boldsymbol{A}\,\mathrm{d}z = -p\,\mathrm{d}V\,, \tag{4.9}$$

and thus

$$W_{v12} = -\int_1^2 p\,\mathrm{d}V\,. \tag{4.10}$$

The minus sign is due to the formal sign convention which states that work transferred to the system, which is connected to a volume reduction, is positive. Equation (4.10) is only valid if the pressure p of the system is in each instance of the change of state a continuous function of volume and equal to the pressure exerted by the environment. Then a small excess or negative pressure of the environment causes either a decrease or an increase of the system volume. Such changes between states, where even the slightest imbalance is sufficient to drive them in either direction, are called reversible. Accordingly, (4.10) is the moving boundary work for reversible changes of state. In real processes a finite excess pressure of the environment is necessary to overcome the internal friction of the system. Such changes in state are irreversible, where the added work is increased by the dissipated part W_{diss12}. The moving boundary work for an irreversible change of state is

$$W_{v12} = -\int_1^2 p\,\mathrm{d}V + W_{diss12}\,. \tag{4.11}$$

The dissipation work is always positive and increases the system energy and causes a different path $p(V)$

between the states than in the reversible case. The determination of the integral in (4.11) requires that p is a unique function of V. Equation (4.11) is, for example, not valid for a system area through which a sound wave travels.

Work can be derived as the product of a generalized force F_k and a generalized displacement dX_k. In real processes the dissipated work has to be added

$$dW = \sum F_k \, dX_k + dW_{diss} \, . \tag{4.12}$$

This equation shows that in irreversible processes ($W_{diss} > 0$) more work has to be done or less work is received than in reversible processes ($W_{diss} = 0$). Table 4.3 includes different forms of work.

Shaft work is work derived from a mass flow through a machine such as compressors, turbines, and jet engines. When a machine increases the pressure of a mass m along the path dz by dp, the shaft work is

$$dW_t = mv \, dp + dW_{diss} \, . \tag{4.13}$$

When kinetic energy and potential energy of the mass flow are also changed, mechanical work is done additionally. The shaft work done along path 1–2 is

$$W_{t12} = \int_1^2 V \, dp + W_{diss12} + W_{m12} \, , \tag{4.14}$$

with W_{m12} is given according to (4.8).

Internal Energy

In addition to any kinetic and potential energy, every system possesses energy stored internally as translational, rotational, and vibrational kinetic energy of the elementary particles. This is called the internal energy U of the system and is an extensive property. The total energy E a system of mass m possesses consists of internal energy, kinetic energy E_{kin}, and potential energy E_{pot}

$$E = U + E_{kin} + E_{pot} \, . \tag{4.15}$$

Heat Transfer

The internal energy of a system can be changed by doing work on it or by adding or removing matter. However, it can also be changed by exposing the system to its environment which has a different temperature. As a consequence, energy is transferred across the system boundary as the system will try to reach thermal equilibrium with the environment. This energy transfer is called heat transfer. Thus heat transfer can generally be defined as that energy a system exchanges with its environment which does not cross the system boundary as work or by accompanying mass transfer. The heat transfer from state 1 to 2 is denoted Q_{12}.

4.3.3 Application to Closed Systems

The heat transfer Q_{12} and work W_{12} to a closed system during the change of state from 1 to 2 cause a change of the system energy E

$$E_2 - E_1 = Q_{12} + W_{12} \, , \tag{4.16}$$

where W_{12} includes all forms of work done on the system. If no mechanical work is done, only the internal energy changes, and according to (4.15), $E = U$ holds. If it is additionally assumed that only moving boundary work is done on the system, (4.16) reads

$$U_2 - U_1 = Q_{12} - \int_1^2 p \, dV + W_{diss12} \, . \tag{4.17}$$

4.3.4 Application to Open Systems

Steady-State Processes

Very often work is done by a fluid flowing steadily through a device. If the work per unit time remains constant, such a process is called a *steady flow process*. Figure 4.2 shows a typical example: a flowing fluid (gas or liquid) of pressure p_1 and temperature T_1 may flow with velocity w_1 into system σ. If machine work is done as shaft work, W_{t12} is supplied at the shaft. Then the fluid flows through a heat exchanger, in which the heat transfer Q_{12} occurs with the environment, and the fluid eventually leaves the system σ with pressure p_2, temperature T_2, and velocity w_2. Tracking the path of a constant mass element Δm through the system σ means that a moving observer would consider the mass element Δm as a closed system, thus this corresponds to the Lagrangian description in fluid mechanics. Therefore, the first law for closed systems (4.16) is valid in this case. The work done on Δm consists of $\Delta m \, p_1 v_1$ to push Δm out of the environment across the system boundary, of the technical work W_{t12}, and of $-\Delta m \, p_2 v_2$ to bring Δm back into the environment. Thus, the work done on the closed system is

$$W_{12} = W_{t12} + \Delta m (p_1 v_1 - p_2 v_2) \, . \tag{4.18}$$

The term $\Delta m \, (p_1 v_1 - p_2 v_2)$ is referred to as the *flow work*. This flow work is the difference between

Table 4.3 Different forms of work. SI units are given in brackets

Form of work	Generalized force	Generalized displacement	Work done
Linear elastic displacement	Force F (N)	Displacement $\mathrm{d}z$ (m)	$\mathrm{d}W = F\,\mathrm{d}z = \sigma\,\mathrm{d}\varepsilon V$ (Nm)
Rotation of a rigid body	Torque M_d (Nm)	Torsion angle $\mathrm{d}\alpha$ (−)	$\mathrm{d}W = M_\mathrm{d}\,\mathrm{d}\alpha$ (Nm)
Moving boundary work	Pressure p (N/m^2)	Volume $\mathrm{d}V$ (m^3)	$\mathrm{d}W_\mathrm{v} = -p\,\mathrm{d}V$ (Nm)
Surface enlargement	Surface tension σ' (N/m)	Area $\mathrm{d}A$ (m^2)	$\mathrm{d}W = \sigma'\,\mathrm{d}A$ (Nm)
Electric work	Voltage U_e (V)	Charge Q_e (C)	$\mathrm{d}W = U_\mathrm{e}\,\mathrm{d}Q_\mathrm{e}$ (Ws) in a linear conductor of resistance R $\mathrm{d}W = U_\mathrm{e} I\,\mathrm{d}t$ $= RI^2\,\mathrm{d}t$ $= (U^2/R)\,\mathrm{d}t$ (Ws)
Magnetic work, in vacuum	Magnetic field strength H_0 (A/m)	Magnetic induction $\mathrm{d}B_0 = \mu_0 H_0$ (Vs/m^2)	$\mathrm{d}W_\mathrm{V} = \mu_0 H_0\,\mathrm{d}H_0$ (Ws/m^3)
Magnetization	Magnetic field strength H (A/m)	Magnetic induction $\mathrm{d}B = \mathrm{d}(\mu_0 H + M)$ (Vs/m^2)	$\mathrm{d}W_\mathrm{V} = H\,\mathrm{d}B$ (Ws/m^3)
Electrical polarization	Electric field strength E (V/m)	Dielectric displacement $\mathrm{d}D = \mathrm{d}(\varepsilon_0 E + P)$ (As/m^2)	$\mathrm{d}W_\mathrm{V} = E\,\mathrm{d}D$ (Ws/m^3)

the shaft work W_{t12} and the work done on the closed system. With this relationship the first law for closed systems, (4.16), becomes

$$E_2 - E_1 = Q_{12} + W_{t12} + \Delta m(p_1 v_1 - p_2 v_2) \quad (4.19)$$

with E according to (4.15). The property enthalpy is defined as

$$H = U + pV \quad \text{or} \quad h = u + pv \quad (4.20)$$

and (4.19) then can be written as

$$0 = Q_{12} + W_{t12} + \Delta m\left(h_1 + \frac{w_1^2}{2} + gz_1\right) - \Delta m\left(h_2 + \frac{w_2^2}{2} + gz_2\right). \quad (4.21)$$

In this form the first law is used for steady flow processes in open systems. Equation (4.21) shows that the sum of all energies entering or leaving the system across the system boundary σ (Fig. 4.2) is zero, because a steady flow process is considered. These energies are in the form of the heat transfer Q_{12}, the shaft work W_{t12}, and the energies $\Delta m(h_1 + w_1^2/2 + gz_1)$ transferred to the system and $\Delta m(h_2 + w_2^2/2 + gz_2)$ transferred from the system with the mass Δm. The differential form of (4.21) reads

$$0 = \mathrm{d}Q + \mathrm{d}W_\mathrm{t} + \mathrm{d}m\left(h_1 + \frac{w_1^2}{2} + gz_1\right) - \mathrm{d}m\left(h_2 + \frac{w_2^2}{2} + gz_2\right).$$

When a continuous process is considered, it is better to use the following form of the balance equation instead of (4.21)

$$0 = \dot{Q} + P + \dot{m}\left(h_1 + \frac{w_1^2}{2} + gz_1\right) - \dot{m}\left(h_2 + \frac{w_2^2}{2} + gz_2\right).$$

In the above equation $\dot{Q} = \mathrm{d}Q/\mathrm{d}\tau$ is the heat transfer rate, $P = \mathrm{d}W_\mathrm{t}/\mathrm{d}\tau$ the shaft power, and $\dot{m}$ the mass flow rate. Changes of kinetic and potential energy in these cases are often negligible, such that (4.21) is simplified to

$$0 = Q_{12} + W_{t12} + H_1 - H_2. \quad (4.22)$$

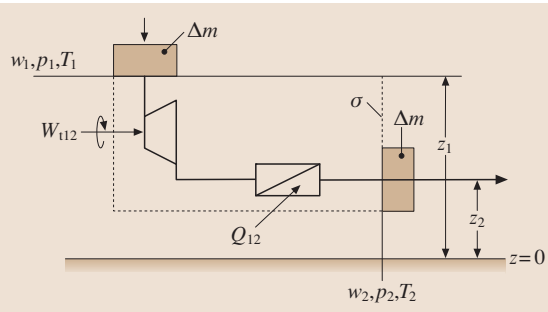


Fig. 4.2 Work for an open system

Special cases of this equation are:

a) Adiabatic changes of state, which typically appear in devices such as compressors, turbines, and jet engines

$$0 = W_{t12} + H_1 - H_2 \,. \tag{4.23}$$

b) Throttling of a flow in an adiabatic tube through a restriction (Fig. 4.3) which causes a pressure reduction. It holds that

$$H_1 = H_2 \tag{4.24}$$

before and after the throttling valve. Thus, the enthalpy remains constant during the throttling, assuming that changes of kinetic and potential energies are negligible.

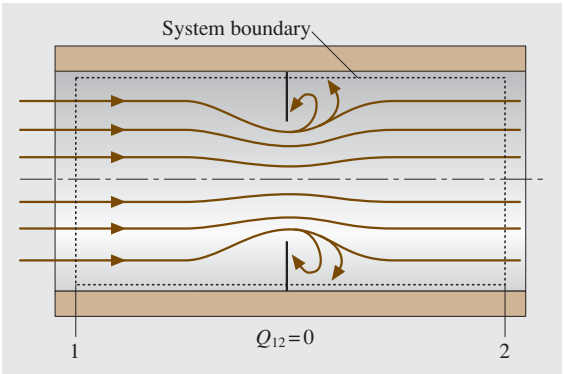


Fig. 4.3 Adiabatic throttling

Transient Processes

Referring to Fig. 4.2, when the mass Δm_1 transferred to the system over a period of time differs from the mass Δm_2 transferred from the system during the same period, the result is mass stored (or loss) in the system. This results in a time-variable internal energy of the system and possibly also time-variable kinetic and potential energies. The energy change of a system during a change of state 1–2 is $E_2 - E_1$. Therefore, (4.21) has to be replaced by the following form of the first law

$$E_2 - E_1 = Q_{12} + W_{t12} + \Delta m_1\left(h_1 + \frac{w_1^2}{2} + g z_1\right) - \Delta m_2\left(h_2 + \frac{w_2^2}{2} + g z_2\right) . \tag{4.25}$$

If the fluid states 1 at the inlet and 2 at the outlet vary in time, it is appropriate to use the differential notation:

$$\mathrm{d}E = \mathrm{d}Q + \mathrm{d}W_t + \mathrm{d}m_1\left(h_1 + \frac{w_1^2}{2} + g z_1\right) - \mathrm{d}m_2\left(h_2 + \frac{w_2^2}{2} + g z_2\right) . \tag{4.26}$$

When investigating the filling and emptying of containers it is usually possible to neglect changes in kinetic and potential energies. Furthermore, often no shaft work is done, thus, (4.26) is reduced to

$$\mathrm{d}U = \mathrm{d}Q + h_1\,\mathrm{d}m_1 - h_2\,\mathrm{d}m_2 \tag{4.27}$$

with the (time-variable) internal energy $U = um$ of the mass stored in the container. It is agreed that $\mathrm{d}m_1$ is mass transferred to and $\mathrm{d}m_2$ mass transferred from the system. If mass is only supplied, $\mathrm{d}m_2$ is equal to zero; if mass is only discharged, $\mathrm{d}m_1$ is equal to zero. For a continuously running process the following form of the balance equation is more suitable than (4.25)

$$\mathrm{d}E/\mathrm{d}\tau = \dot{Q} + P + \dot{m}_1\left(h_1 + \frac{w_1^2}{2} + g z_1\right) - \dot{m}_2\left(h_2 + \frac{w_2^2}{2} + g z_2\right) . \tag{4.28}$$

4.4 Second Law of Thermodynamics

4.4.1 The Principle of Irreversibility

When two systems A and B are exposed to each other, energy exchange processes take place and a new equilibrium state is reached after a sufficient period of time. As an example, a system A may be in contact with a system B that has a different temperature. In the final state both systems will have the same temperature and equilibrium will have been reached. Until equilibrium has been reached a continuous series of nonequilibrium states will be passed. It is from common experience that this process has a natural direction (e.g., heat transfer from hot to cold) and does not proceed in the reverse direction independently, i. e., without exchange with the environment. Such processes are referred to as being irreversible. Exchange processes, which pass through nonequilibrium states, are in principle irreversible. On the other hand, a process that consists of a continu-

ous series of equilibrium states is reversible. This may be exemplified by the frictionless, adiabatic compression of a gas. It is possible to transfer moving boundary work to the system *gas* by exerting a force, for example, an excess pressure of the environment, on the system boundary. If this force is increased very slowly, the volume of the gas decreases and the temperature increases, whereas the gas is at any time in an equilibrium state. If the force is slowly reduced to zero, the gas returns to its initial state; thus, this process is reversible. Reversible processes are idealized borderline cases of real processes and do not occur in nature. All natural processes are irreversible, because a finite force is necessary to initiate a process, e.g., a finite force to move a body against friction or a finite temperature difference for heat transfer. These facts known from experience lead to the following formulations of the second law:

- All natural processes are irreversible.
- All processes including friction are irreversible.
- Heat transfer does not independently occur from a body of lower to a body of higher temperature.

Independently in this connection means that it is not possible to carry out the mentioned process without causing effects on nature. Beside these examples, further formulations valid for other special processes exist.

4.4.2 General Formulation

The mathematical formulation of the second law is realized by introducing the term *entropy* as another property of a system. The practicality of this property can be shown by using the example of heat transfer between a system and its environment. According to the first law, a system can exchange energy by work and by heat transfer with its environment. The supply of work causes a change of the internal energy such that, for example, the system's volume is changed at the expense of the environment's volume. Consequently, $U = U(V, \ldots)$. The volume is an exchange variable. It is an extensive property, which is *exchanged* between the system and environment. It is also possible to look upon the heat transfer between a system and its environment as an exchange of an extensive property. In this way, only the existence of such a property is postulated. Its introduction is solely justified by the fact that all statements derived from it correspond with experience. This new extensive property is called entropy and denoted with S. Consequently, $U = U(V, S, \ldots)$. If only moving boundary work occurs and heat transfer occurs, $U = U(V, S)$. Differentiation leads to the *Gibbs equation*

$$\mathrm{d}U = T\,\mathrm{d}S - p\,\mathrm{d}V \tag{4.29}$$

with the thermodynamic temperature

$$T = (\partial U/\partial S)_\mathrm{V} \tag{4.30}$$

and the pressure

$$p = -(\partial U/\partial V)_\mathrm{S}\,. \tag{4.31}$$

A relationship equivalent to (4.29) is derived by eliminating U and replacing it by enthalpy $H = U + pV$ such that

$$\mathrm{d}H = T\,\mathrm{d}S + V\,\mathrm{d}p\,. \tag{4.32}$$

It can be shown that the thermodynamic temperature is identical to the temperature measured by a gas thermometer (Sect. 4.2.3). From examination of the characteristics of entropy it follows that in an isolated system, which is initially in nonequilibrium (for example, because of a nonuniform temperature distribution) and then approaches equilibrium, the entropy always increases. In the borderline case of equilibrium a maximum of entropy is reached. The internal entropy increase is denoted by $\mathrm{d}S_\mathrm{gen}$. For the considered case of an isolated system it holds that

$$\mathrm{d}S = \mathrm{d}S_\mathrm{gen}$$

with $\mathrm{d}S_\mathrm{gen} > 0$. If a system is not isolated, entropy is also changed by $\mathrm{d}S_\mathrm{Q}$ due to heat transfer (with the environment) and by $\mathrm{d}S_\mathrm{m}$ because of mass transfer with the environment. However, energy transfer by work with the environment does not change the system entropy. Thus, it holds generally that

$$\mathrm{d}S = \mathrm{d}S_\mathrm{Q} + \mathrm{d}S_\mathrm{m} + \mathrm{d}S_\mathrm{gen}\,. \tag{4.33}$$

The formulation for the time-variable system entropy $\dot{S} = \mathrm{d}S/\mathrm{d}\tau$ reads

$$\dot{S} = \dot{S}_\mathrm{Q} + \dot{S}_m + \dot{S}_\mathrm{gen} \tag{4.34}$$

with $\dot{S}_\mathrm{gen}$ being the entropy generation rate caused by internal irreversibilities, and $\dot{S}_\mathrm{Q} + \dot{S}_\mathrm{m}$ is called the entropy flow. These values, which are exchanged across the system boundary, are combined to

$$\dot{S}_\mathrm{fl} = \dot{S}_\mathrm{Q} + \dot{S}_\mathrm{m}\,. \tag{4.35}$$

The rate of change of the system entropy S consists, thus, of the entropy flow $\dot{S}_\mathrm{fl}$ and entropy generation $\dot{S}_\mathrm{gen}$

$$\dot{S} = \dot{S}_\mathrm{fl} + \dot{S}_\mathrm{gen}\,. \tag{4.36}$$

For the entropy generation it holds that

$$\begin{aligned} \dot{S}_{\mathrm{gen}} &= 0 \quad \text{for reversible processes,} \\ \dot{S}_{\mathrm{gen}} &> 0 \quad \text{for irreversible processes,} \\ \dot{S}_{\mathrm{gen}} &< 0 \quad \text{for impossible processes.} \end{aligned} \tag{4.37}$$

4.4.3 Special Formulations

Adiabatic, Closed Systems

Since $\dot{S}_{\mathrm{Q}} = 0$ for adiabatic systems and $\dot{S}_{\mathrm{m}} = 0$ for closed systems, it follows that $\dot{S} = \dot{S}_{\mathrm{gen}}$. Thus, the entropy of an adiabatic, closed system can never decrease. It can only increase during an irreversible process or remain constant during a reversible process. If an adiabatic, closed system consists of α subsystems, then it holds for the sum of entropy changes ΔS^{α} of the subsystems that

$$\sum_{\alpha} \Delta S^{\alpha} \geq 0 \,. \tag{4.38}$$

With $\mathrm{d}S = \mathrm{d}S_{\mathrm{gen}}$, (4.29) reads for an adiabatic, closed system

$$\mathrm{d}U = T\,\mathrm{d}S_{\mathrm{gen}} - p\,\mathrm{d}V \,.$$

On the other hand it follows from the first law according to (4.17)

$$\mathrm{d}W_{\mathrm{diss}} = T\,\mathrm{d}S_{\mathrm{gen}} = \mathrm{d}\Psi \tag{4.39}$$

or

$$W_{\mathrm{diss}12} = TS_{\mathrm{gen}12} = \Psi_{12} \,, \tag{4.40}$$

where Ψ_{12} is called the dissipated energy during the change in state 1–2. The dissipated energy is always positive. This statement is not only true for adiabatic systems but also for all general cases, because, according to definition, the entropy generation is the fraction of entropy change, which arises when the system is adiabatic and closed and therefore $\dot{S}_{\mathrm{fl}} = 0$ holds.

Systems with Heat Transfer

For closed systems with heat transfer (4.29) becomes

$$\begin{aligned} \mathrm{d}U &= T\,\mathrm{d}S_{\mathrm{Q}} + T\,\mathrm{d}S_{\mathrm{gen}} - p\,\mathrm{d}V \\ &= T\,\mathrm{d}S_{\mathrm{Q}} + \mathrm{d}W_{\mathrm{diss}} - p\,\mathrm{d}V \,. \end{aligned} \tag{4.41}$$

A comparison with the first law, (4.17), results in

$$\mathrm{d}Q = T\,\mathrm{d}S_{\mathrm{Q}} \,. \tag{4.42}$$

Thus, heat transfer is energy transfer, which together with entropy crosses the system boundary, whereas work is exchanged without entropy exchange. Adding the always positive term $T\,\mathrm{d}S_{\mathrm{gen}}$ to the right-hand side of (4.42) leads to the *Clausius inequality*

$$\mathrm{d}Q \leq T\,\mathrm{d}S \quad \text{or} \quad \Delta S \geq \int_1^2 \frac{\mathrm{d}Q}{T} \,. \tag{4.43}$$

In irreversible processes the entropy change is larger than the integral over all $\mathrm{d}Q/T$; the equals sign is only valid for the reversible case. For open systems with heat addition, $\mathrm{d}S_{\mathrm{Q}}$ in (4.41) has to be replaced by $\mathrm{d}S_{\mathrm{fl}} = \mathrm{d}S_{\mathrm{Q}} + \mathrm{d}S_{\mathrm{m}}$.

4.5 Exergy and Anergy

According to the first law, the energy of an isolated system is constant. As it is possible to transform every nonisolated system into an isolated one by adding the environment, it is always possible to define a system in which the energy remains constant during a thermodynamic process. Thus, a loss of energy is not possible, and energy is only converted in a thermodynamic process. How much of the energy stored in a system is converted depends on the state of the environment. If it is in equilibrium with the system, no energy is converted. The larger the difference from equilibrium, the more energy of the system can be converted and thus the greater the potential to perform work.

Many thermodynamic processes take place in the Earth's atmosphere, which is the environment of most thermodynamic systems. In comparison to the much smaller thermodynamic systems, the Earth's atmosphere can be considered as an infinitely large system, in which the intensive properties pressure, temperature, and composition do not change during a process (as long as daily and seasonal variations of the intensive properties are neglected). In many engineering processes work is obtained by bringing a system with a given initial state into equilibrium with the environment. The maximum work is obtained when all changes of state are reversible.

The maximum work that could be obtained by establishing equilibrium with the environment is called the exergy W_{ex}.

4.5.1 Exergy of a Closed System

In order to calculate the exergy of a system at state 1, a process is considered that brings the system reversibly into thermal and mechanical equilibrium with its environment. Equilibrium exists if the temperature of the system at the final state 2 is equal to the temperature of the environment, i. e., $T_2 = T_{env}$, and if the pressure of the system in state 2 is equal to the pressure of the environment, i. e., $p_2 = p_{env}$. Neglecting the kinetic and potential energy of the system, the first law according to (4.16) reads

$$U_2 - U_1 = Q_{12} + W_{12} \,. \tag{4.44}$$

To execute the entire process reversibly, it is necessary to bring the system to the environment temperature through a reversible, adiabatic change of state. Then heat transfer has to occur reversibly at the constant temperature T_{env}.

From the second law, (4.42), it follows for the heat transfer that

$$Q_{12} = T_{env}(S_2 - S_1) \,. \tag{4.45}$$

The work W_{12} done on the system consists of the maximum useful work and the moving boundary work $-p_{env}(V_2 - V_1)$, which is necessary to overcome the pressure of the environment. The maximum useful work is the exergy W_{ex}, thus it follows that

$$W_{12} = W_{ex} - p_{env}(V_2 - V_1) \,. \tag{4.46}$$

Inserting (4.45) and (4.46) into (4.44) gives

$$U_2 - U_1 = T_{env}(S_2 - S_1) + W_{ex} - p_{env}(V_2 - V_1) \,. \tag{4.47}$$

In state 2 the system is in equilibrium with the environment, denoted by the index 'env'. Thus, the exergy of the closed system is

$$-W_{ex} = U_1 - U_{env} - T_{env}(S_1 - S_{env}) + p_{env}(V_1 - V_{env}) \,. \tag{4.48}$$

For a constant-volume system it holds that $V_1 = V_{env}$ and the last term is cancelled. If the initial state of the system already is in equilibrium with the environment (state 1 = state env), according to (4.48) no work can be obtained.

Thus it holds that the internal energy of the environment cannot be transformed into exergy. Consequently, the enormous energies stored in the atmosphere surrounding us cannot be used to power vehicles.

4.5.2 Exergy of an Open System

The maximum shaft work, or the exergy from a mass flow, is obtained when the mass flow is brought reversibly into equilibrium with the environment by performing work and heat transfer with the environment. Neglecting changes of kinetic and potential energies, the first law for steady flow processes in open systems, (4.22), reads

$$-W_{ex} = H_1 - H_{env} - T_{env}(S_1 - S_{env}) \,. \tag{4.49}$$

This means that only a part of the enthalpy, H_1 reduced by $H_{env} + T_{env}(S_1 - S_{env})$, is transformed into shaft work. If the heat transfer from the environment to the mass flow, $T_{env}(S_1 - S_{env})$, is negative then the exergy exceeds the change in enthalpy by the fraction of this added heat.

4.5.3 Exergy and Heat Transfer

Figure 4.4 shows a device which is used to transform the heat transfer Q_{12} from an energy storage of temperature T into work W_{12}. The heat transfer Q_{env12}, which cannot be transformed into work, is rejected to the environment. The maximum shaft work is obtained if all changes in state are reversible. This maximum shaft work is equal to the exergy of the heat transfer. All changes in state are reversible if

$$\int_1^2 \frac{\mathrm{d}Q}{T} + \int_1^2 \frac{\mathrm{d}Q_{env}}{T_{env}} = 0$$

with $\mathrm{d}Q + \mathrm{d}Q_{env} + \mathrm{d}W_{ex} = 0$ according to the first law. The resulting exergy of the heat transfer to machines and apparatuses is

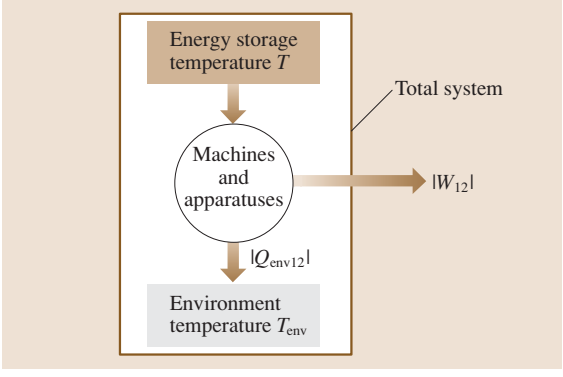


Fig. 4.4 Conversion of heat into work

$$-W_{ex} = \int_1^2 \left(1 - \frac{T_{env}}{T}\right) dQ \tag{4.50}$$

or in differential notation

$$-dW_{ex} = \left(1 - \frac{T_{env}}{T}\right) dQ\,. \tag{4.51}$$

In a reversible process only the fraction of the supplied heat transfer multiplied with the so-called Carnot factor $1-(T_{env}/T)$ can be transformed into work. The fraction $dQ_{env} = -T_{env}(dQ/T)$ has to be transferred to the environment and it is impossible to obtain it as work. This shows additionally that the heat transfer, which is available at ambient temperature, can not be transformed into exergy.

4.5.4 Anergy

Energy that cannot be converted into exergy W_{ex} is called anergy B. Each amount of energy consists of exergy W_{ex} and anergy B, i. e.,

$$E = W_{ex} + B\,. \tag{4.52}$$

Thus it holds that:

- For a closed system according to (4.48) with $E = U_1$

$$B = U_{env} + T_{env}(S_1 - S_{env}) - p_{env}(V_1 - V_{env}) \tag{4.53}$$

- For an open system according to (4.49) with $E = H_1$

$$B = H_{env} + T_{env}(S_1 - S_{env}) \tag{4.54}$$

- For heat transfer according to (4.51) with $dE = dQ$

$$B = \int_1^2 \frac{T_{env}}{T}\, dQ \tag{4.55}$$

4.5.5 Exergy Losses

The energy dissipated in a process is not lost completely. It increases the entropy, and because of $U(S, V)$, also the internal energy of a system. It is possible to think of the dissipated energy as heat transfer in a substitutional process, which is transferred from the outside ($d\Psi = dQ$) and causes the same entropy increase as in the irreversible process. Since the heat transfer dQ, (4.51), is partly transformable into work, the fraction

$$-dW_{ex} = \left(1 - \frac{T_{env}}{T}\right) d\Psi \tag{4.56}$$

of the dissipated energy $d\Psi$ can also be obtained as work (exergy). The remaining fraction $T_{env}\, d\Psi/T$ of the dissipated energy has to be transferred to the environment as heat transfer and is not transformable into work. This exergy loss is equal to the anergy of the dissipated energy and is, according to (4.55), given by

$$W_{loss12} = \int_1^2 \frac{T_{env}}{T}\, d\Psi = \int_1^2 T_{env}\, dS_i\,. \tag{4.57}$$

For a process in a closed, adiabatic system it holds that $dS_i = dS$ and thus

$$W_{loss12} = \int_1^2 T_{env}\, dS = T_{env}(S_2 - S_1)\,. \tag{4.58}$$

In contrast to energy, exergy does not follow a conservation equation. The exergy transferred to a system at steady state is equal to the sum of the exergy transferred from the system plus exergy losses. The thermodynamic effect of losses caused by irreversibilities is more unfavorable for lower temperatures T at which the process takes place; see (4.57).

4.6 Thermodynamics of Substances

In order to utilize the primary laws of thermodynamics, which are generally set up for any arbitrary substance, and to calculate exergies and anergies, it is necessary to determine actual numerical values for the properties U, H, S, p, V, and T. From these U, H, and S typically are called caloric, where p, V, and T are thermal properties. The relationships between properties depend on the corresponding substance. Equations that specify the relationships between the properties p, V, and T are called *equations of state*.

4.6.1 Thermal Properties of Gases and Vapors

An equation of state for pure substances is of the form

$$F(p, v, T) = 0 \tag{4.59}$$

or $p = p(v, T)$, $v = v(p, T)$, and $T = T(p, v)$. For calculations equations of state of the form $v = v(p, T)$ are preferred, as the pressure and temperature are usually the independent variables used to describe a system.

Table 4.4 Critical data of some substances, ordered according to the critical temperature (after [4.4–6])

	Symbol	M (kg/kmol)	P_{cr} (bar)	T_{cr} (K)	v_{cr} (dm^3/kg)
Mercury	Hg	200.59	1490	1765	0.213
Aniline	C_6H_7N	93.1283	53.1	698.7	2.941
Water	H_2O	18.0153	220.55	647.13	3.11
Benzene	C_6H_6	78.1136	48.98	562.1	3.311
Ethyl alcohol	C_2H_5OH	46.0690	61.37	513.9	3.623
Diethyl ether	$C_4H_{10}O$	74.1228	36.42	466.7	3.774
Ethyl chloride	C_2H_5Cl	64.5147	52.7	460.4	2.994
Sulfur dioxide	SO_2	64.0588	78.84	430.7	1.901
Methyl chloride	CH_3Cl	50.4878	66.79	416.3	2.755
Ammonia	NH_3	17.0305	113.5	405.5	4.255
Hydrogen chloride	HCl	36.4609	83.1	324.7	2.222
Nitrous oxide	N_2O	44.0128	72.4	309.6	2.212
Acetylene	C_2H_2	26.0379	61.39	308.3	4.329
Ethane	C_2H_6	30.0696	48.72	305.3	4.926
Carbon dioxide	CO_2	44.0098	73.77	304.1	2.139
Ethylene	C_2H_4	28.0528	50.39	282.3	4.651
Methane	CH_4	16.0428	45.95	190.6	6.173
Nitrous monoxide	NO	30.0061	65	180	1.901
Oxygen	O_2	31.999	50.43	154.6	2.294
Argon	Ar	39.948	48.65	150.7	1.873
Carbon monoxide	CO	28.0104	34.98	132.9	3.322
Air	–	28.953	37.66	132.5	3.195
Nitrogen	N_2	28.0134	33.9	126.2	3.195
Hydrogen	H_2	2.0159	12.97	33.2	32.26
Helium-4	He	4.0026	2.27	5.19	14.29

Ideal Gases

A particularly simple equation of state is that for ideal gases

$$pV = mRT \quad \text{or} \quad pv = RT, \tag{4.60}$$

which relates the absolute pressure p, the volume V, the specific volume v, the individual gas constant R, and the thermodynamic temperature T. A gas is assumed to behave as an ideal gas only when the pressure is sufficiently low compared to the critical pressure p_{cr} of the substance, i. e., $p/p_{cr} \to 0$.

Gas Constant and Avogadro's Law

As a measure of the amount of a given system, the *mole* is defined with the unit symbol mol. The amount of a substance is 1 mol when it contains as many identical elementary entities (i. e., molecules, atoms, elementary particles) as there are atoms in exactly 12 g of pure carbon-12.

The number of particles of the same type contained in a mole is called Avogadro's number (in German literature the number is sometimes referred to as Loschmidt's number). It has the numerical value

$$N_A = (6.02214199 \pm 4.7 \times 10^{-7}) \times 10^{26} \text{ /kmol} . \tag{4.61}$$

The mass of a mole (N_A particles of the same type) is a specific quantity for each substance and is referred to as the molar mass (see tab003-9 for values), which is given by

$$M = m/n \tag{4.62}$$

(SI unit kg/kmol, m mass in kg, n molar amount in kmol). According to Avogadro (1811), ideal gases contain an equal number of molecules at the same pressure and at the same temperature occupying equal spaces.

After introducing the molar mass into the equation of state for ideal gases, (4.60), it follows that $pV/nT = MR$ has a fixed value for all gases

$$MR = R_u, \tag{4.63}$$

where R_u is called the universal gas constant, and is a fundamental constant with the numerical value

$$R_u = 8.314472 \pm 1.5 \times 10^{-5}\ \text{kJ/kmol K}\,. \tag{4.64}$$

Incorporating R_u, the equation of state for ideal gases reads

$$pV = nR_uT\,. \tag{4.65}$$

Example 4.6: A gas bottle of volume $V_1 = 200\,\text{l}$ contains hydrogen at $p_1 = 120\,\text{bar}$ and $t_1 = 10\,°\text{C}$. What space is occupied by the hydrogen at $p_2 = 1\,\text{bar}$ and $t_2 = 0\,°\text{C}$, if the hydrogen is assumed to behave as an ideal gas? According to (4.65), $p_1V_1 = nRT_1$ and $p_2V_2 = nRT_2$; thus

$$V_2 = \frac{p_1T_2}{p_2T_1}V_1 = \frac{120\,\text{bar} \times 273.15\,\text{K}}{1\,\text{bar} \times 283.15\,\text{K}} 0.2\,\text{m}^3 = 23.15\,\text{m}^3\,. \tag{4.66}$$

Real Gases

The ideal gas equation of state is valid for real gases and vapors only as a limiting law at sufficiently low pressures. The deviation of the behavior of gaseous water from the ideal gas equation of state is shown in Fig. 4.5, in which pv/RT is displayed against t for different pressures. The real gas factor Z, where $Z = pv/RT$, is equal to 1 for ideal gases, but deviates from this value for real gases. For air between 0 °C and 200 °C, and for hydrogen between −15 °C and 200 °C, the deviations of Z from ideal gas behavior are approximately 1% at a pressure of 20 bar. At atmospheric pressure the deviations from the ideal gas law are negligible for nearly all gases. For cases where significant deviation is found from ideal gas behavior, different equations of state are established to describe the behavior of real gases. In one of the simplest forms, the real gas factor Z, a series of additive correction terms are used to modify its value from unity for the ideal gas case

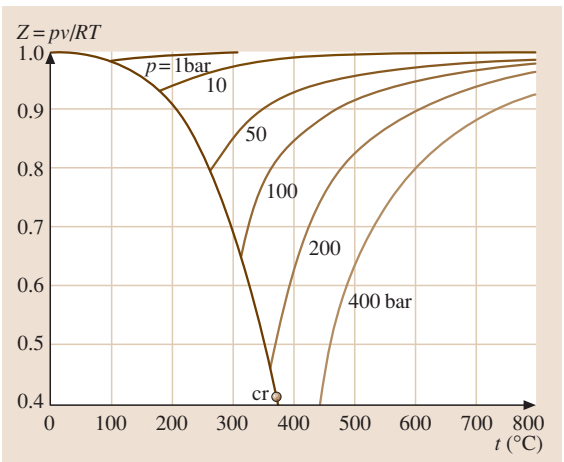


Fig. 4.5 The real gas factor of water vapor

$$Z = \frac{pv}{RT} = 1 + \frac{B(T)}{v} + \frac{C(T)}{v^2} + \frac{D(T)}{v^3}\,, \tag{4.67}$$

where B is called the second, C the third, and D the fourth virial coefficients. A compilation of second virial coefficients for many gases is provided by reference charts [4.7, 8]. The virial equation with two or three virial coefficients is only valid at moderate pressures. A balanced compromise between computational effort and achievable accuracy is given by the equation of state by Benedict–Webb–Rubin [4.9] for denser gases, which reads

$$Z = 1 + \frac{B(T)}{v} + \frac{C(T)}{v^2} + \frac{a\alpha}{v^5RT} + \frac{c}{v^3RT^2}\left(1 + \frac{\gamma}{v^2}\right)\exp\left(-\frac{\gamma}{v^2}\right)\,, \tag{4.68}$$

with

$$B(T) = B_0 - \frac{A_0}{RT} - \frac{C_0}{RT^3} \quad \text{and} \quad C(T) = b - \frac{a}{RT}\,.$$

The equation contains the eight constants A_0, B_0, C_0, a, b, c, α, and γ, which are available for many substances [4.9]. Highly exact equations of state are needed for the working substances water [4.10], air [4.11], and refrigerants [4.12] used in heat engines and refrigerators. The equations for these substances are more elaborate, contain more constants, and computer software is typically needed to utilize them.

Vapors

Vapors are gases which are close to saturation conditions and to condensation. A vapor is called *saturated* if the slightest temperature reduction leads to condensation, and *superheated* if a finite temperature reduction is necessary to obtain condensation. If heat is transferred to a liquid at constant pressure, the temperature of the liquid rises. When a certain temperature is reached,

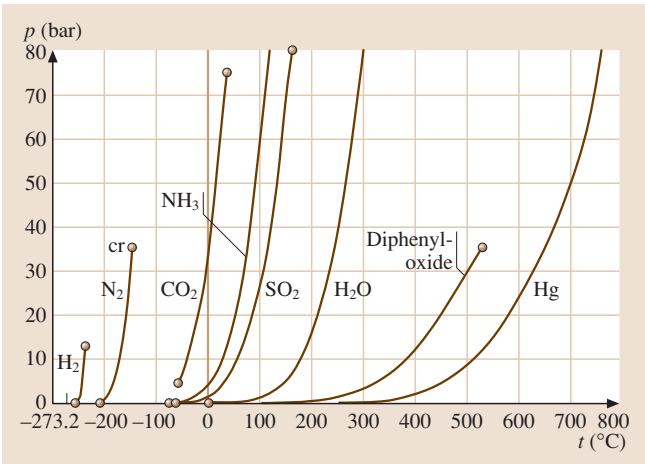


Fig. 4.6 Liquid–vapor saturation curves of some substances

vapor at the same temperature begins to be generated, where the vapor and liquid are in equilibrium. This state is called the saturation state. It is characterized by corresponding values for *saturation temperature* and *saturation pressure*. Their interdependence is described by the *liquid–vapor saturation curve* in Fig. 4.6. It starts at the triple point and ends at the critical point (cr) of a substance. Above the critical state p_{cr}, T_{cr}, vapor and liquid are no longer separated by a clear boundary but merge continuously (see Table 4.4). As with the triple point, at which vapor, liquid and solid phases of a substance are in equilibrium, every substance also has a characteristic critical point.

The vapor pressure of many substances is well approximated between the triple point and the boiling

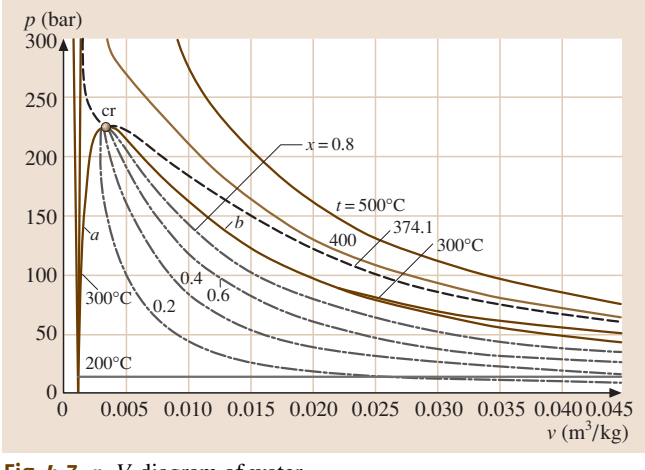


Fig. 4.7 *p–V* diagram of water

Table 4.5 Antoine equation ($\log_{10} p = A - \frac{B}{C+t}$, p in hPa, t in °C), constants for some substances (after [4.13])

Sustance	*A*	*B*	*C*
Methane	6.82051	405.42	267.777
Ethane	6.95942	663.70	256.470
Propane	6.92888	803.81	246.99
Butane	6.93386	935.86	238.73
Isobutene	7.03538	946.35	246.68
Pentane	7.00122	1075.78	233.205
Isopentane	6.95805	1040.73	235.445
Neopentane	6.72917	883.42	227.780
Hexane	6.99514	1168.72	224.210
Heptane	7.01875	1264.37	216.636
Octane	7.03430	1349.82	209.385
Cyclopentane	7.01166	1124.162	231.361
Methylcyclopentane	6.98773	1186.059	226.042
Cyclohexane	6.96620	1201.531	222.647
Methylcyclohexane	6.94790	1270.763	221.416
Ethylene	6.87246	585.00	255.00
Propylene	6.94450	785.00	247.00
Butylene-(1)	6.96780	926.10	240.00
Butylene-(2) cis	6.99416	960.100	237.000
Butylene-(2) trans	6.99442	960.80	240.00
Isobutylene	6.96624	923.200	240.000
Pentylene-(1)	6.97140	1044.895	233.516
Hexylene-(1)	6.99063	1152.971	225.849
Propadiene	5.8386	458.06	196.07
Butadiene-(1,3)	6.97489	930.546	238.854
2-Methylbutadiene	7.01054	1071.578	233.513
Benzene	7.03055	1211.033	220.790
Toluene	7.07954	1344.800	219.482
Ethylbenzene	7.08209	1424.255	213.206
m-Xylene	7.13398	1462.266	215.105
p-Xylene	7.11542	1453.430	215.307
Isopropylbenzene	7.06156	1460.793	207.777
Water (90–100 °C)	8.0732991	1656.390	226.86

point at atmospheric pressure by the Antoine equation

$$\ln p = A - \frac{B}{C+T}\,, \tag{4.69}$$

in which A, B, and C are the *Antoine coefficients* that vary from substance to substance (see Table 4.5). If superheated vapor is compressed at constant temperature by reducing the volume, the pressure increases similar to an ideal gas almost like a hyperbola, e.g., see the 300 °C isotherm in Fig. 4.7. Condensation starts as soon as the saturation pressure is reached, and the

volume is reduced without a pressure increase until all vapor is condensed. Any further volume reduction causes a considerable pressure increase. The band of curves in Fig. 4.7 is, as a graphical description of an equation of state, characteristic for many substances. Connecting the specific volumes of the liquid at saturation temperature before evaporation and of the saturated vapor, v' and v'', results in two curves a and b, called the *saturated liquid line* and the *saturated vapor line*, which meet at the critical point. With the steam quality x, defined as the mass of the saturated vapor m'' related to the total mass of saturated vapor m'' and saturated liquid m', and the specific volumes v' of the saturated liquid and v'' of the saturated vapor, it holds for wet steam that

$$v = xv'' + (1-x)v' \,. \tag{4.70}$$

Lines of constant x are shown in Fig. 4.7.

Example 4.7: 1000 kg saturated wet steam at 121 bar is in a vessel of 2 m^3 volume. How is the total mass distributed between liquid and vapor? An interpolation of the values in the saturated water table (Table 4.6) leads to the specific volumes $v' = 0.001530\,m^3/kg$ of saturated liquid and $v'' = 0.01410\,m^3/kg$ of saturated vapor at 121 bar. The average specific volume $v = V/m$ is $v = 2\,m^3/1000\,kg = 0.002\,m^3/kg$. Equation (4.70) gives

$$\begin{aligned} x &= (v-v')/(v''-v') \\ &= (0.002-0.001530)/(0.01410-0.001530) \\ &= 0.03739 \\ &= m''/m \,, \end{aligned}$$

and thus

$$\begin{aligned} m'' &= 1000 \times 0.03739\,kg = 37.39\,kg \,, \\ m' &= (1000-37.39)\,kg = 962.61\,kg \,. \end{aligned}$$

It also is possible to display the equation of state as a surface in space with coordinates p, v, and t (Fig. 4.8).

The two-dimensional diagrams in Figs. 4.6 and 4.7 are projections of this three-dimensional surface onto the respective planes.

4.6.2 Caloric Properties of Gases and Vapors

Ideal Gases

The internal energy of ideal gases depends only on temperature, $u = u(T)$, and thus also the enthalpy $h = u + pv = u + RT$ is solely a function of temperature, $h = h(T)$. The derivatives of u and h with respect to temperature are called *specific heats*. The specific heats are also functions of temperature and increase with temperature (see Table 4.8 for values of air).

$$du/dT = c_v \tag{4.71}$$

is the specific heat at constant volume and

$$dh/dT = c_p \tag{4.72}$$

the specific heat at constant pressure. The derivative of $h - u = RT$ is

$$c_p - c_v = R \,. \tag{4.73}$$

The difference of the molar specific heats $\overline{C}_p = Mc_p$ and $\overline{C}_v = Mc_v$ is equal to the universal gas constant

$$\overline{C}_p - \overline{C}_v = R_u \,.$$

The specific heat ratio $\varkappa = c_p/c_v$ plays an important role in reversible, adiabatic changes of state and is hence also called an *adiabatic exponent*.

For monatomic gases the specific heat ratio is fairly accurately $\varkappa = 1.66$, for diatomic gases $\varkappa = 1.40$, and

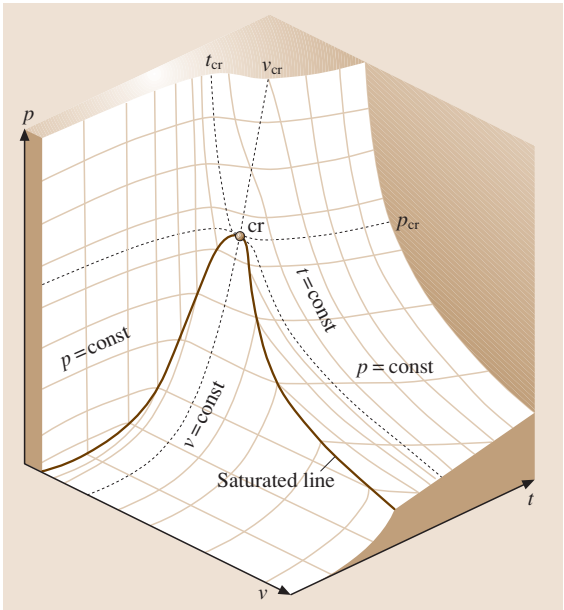


Fig. 4.8 Area of states for water

Table 4.6 Saturated water temperature table (after [4.10])

t (°C)	p (bar)	v' (m^3/kg)	v'' (m^3/kg)	h' (kJ/kg)	h'' (kJ/kg)	Δh_v (kJ/kg)	s' (kJ/(kgK))	s'' (kJ/(kgK))
0.01	0.006117	0.001000	205.998	0.00	2500.91	2500.91	0.0000	9.1555
2	0.007060	0.001000	179.764	8.39	2504.57	2496.17	0.0306	9.1027
4	0.008135	0.001000	157.121	16.81	2508.24	2491.42	0.0611	9.0506
6	0.009354	0.001000	137.638	25.22	2511.91	2486.68	0.0913	8.9994
8	0.010730	0.001000	120.834	33.63	2515.57	2481.94	0.1213	8.9492
10	0.012282	0.001000	106.309	42.02	2519.23	2477.21	0.1511	8.8998
12	0.014028	0.001001	93.724	50.41	2522.89	2472.48	0.1806	8.8514
14	0.015989	0.001001	82.798	58.79	2526.54	2467.75	0.2099	8.8038
16	0.018188	0.001001	73.292	67.17	2530.19	2463.01	0.2390	8.7571
18	0.020647	0.001001	65.003	75.55	2533.83	2458.28	0.2678	8.7112
20	0.023392	0.001002	57.762	83.92	2537.47	2453.55	0.2965	8.6661
22	0.026452	0.001002	51.422	92.29	2541.10	2448.81	0.3250	8.6218
24	0.029856	0.001003	45.863	100.66	2544.73	2444.08	0.3532	8.5783
26	0.033637	0.001003	40.977	109.02	2548.35	2439.33	0.3813	8.5355
28	0.037828	0.001004	36.675	117.38	2551.97	2434.59	0.4091	8.4934
30	0.042467	0.001004	32.882	125.75	2555.58	2429.84	0.4368	8.4521
32	0.047593	0.001005	29.529	134.11	2559.19	2425.08	0.4643	8.4115
34	0.053247	0.001006	26.562	142.47	2562.79	2420.32	0.4916	8.3715
36	0.059475	0.001006	23.932	150.82	2566.38	2415.56	0.5187	8.3323
38	0.066324	0.001007	21.595	159.18	2569.96	2410.78	0.5457	8.2936
40	0.073844	0.001008	19.517	167.54	2573.54	2406.00	0.5724	8.2557
42	0.082090	0.001009	17.665	175.90	2577.11	2401.21	0.5990	8.2183
44	0.091118	0.001009	16.013	184.26	2580.67	2396.42	0.6255	8.1816
46	0.10099	0.001010	14.535	192.62	2584.23	2391.61	0.6517	8.1454
48	0.11176	0.001011	13.213	200.98	2587.77	2386.80	0.6778	8.1099
50	0.12351	0.001012	12.028	209.34	2591.31	2381.97	0.7038	8.0749
52	0.13631	0.001013	10.964	217.70	2594.84	2377.14	0.7296	8.0405
54	0.15022	0.001014	10.007	226.06	2598.35	2372.30	0.7552	8.0066
56	0.16532	0.001015	9.145	234.42	2601.86	2367.44	0.7807	7.9733
58	0.18171	0.001016	8.369	242.79	2605.36	2362.57	0.8060	7.9405
60	0.19946	0.001017	7.668	251.15	2608.85	2357.69	0.8312	7.9082
62	0.21866	0.001018	7.034	259.52	2612.32	2352.80	0.8563	7.8764
64	0.23942	0.001019	6.460	267.89	2615.78	2347.89	0.8811	7.8451
66	0.26183	0.001020	5.940	276.27	2619.23	2342.97	0.9059	7.8142
68	0.28599	0.001022	5.468	284.64	2622.67	2338.03	0.9305	7.7839
70	0.31201	0.001023	5.040	293.02	2626.10	2333.08	0.9550	7.7540
72	0.34000	0.001024	4.650	301.40	2629.51	2328.11	0.9793	7.7245
74	0.37009	0.001025	4.295	309.78	2632.91	2323.13	1.0035	7.6955
76	0.40239	0.001026	3.971	318.17	2636.29	2318.13	1.0276	7.6669
78	0.43703	0.001028	3.675	326.56	2639.66	2313.11	1.0516	7.6388
80	0.47415	0.001029	3.405	334.95	2643.01	2308.07	1.0754	7.6110
82	0.51387	0.001030	3.158	343.34	2646.35	2303.01	1.0991	7.5837
84	0.55636	0.001032	2.932	351.74	2649.67	2297.93	1.1227	7.5567
86	0.60174	0.001033	2.724	360.15	2652.98	2292.83	1.1461	7.5301

Table 4.6 (cont.)

t (°C)	p (bar)	v' (m^3/kg)	v'' (m^3/kg)	h' (kJ/kg)	h'' (kJ/kg)	Δh_v (kJ/kg)	s' (kJ/(kgK))	s'' (kJ/(kgK))
88	0.65017	0.001035	2.534	368.56	2656.26	2287.70	1.1694	7.5039
90	0.70182	0.001036	2.359	376.97	2659.53	2282.56	1.1927	7.4781
92	0.75685	0.001037	2.198	385.38	2662.78	2277.39	1.2158	7.4526
94	0.81542	0.001039	2.050	393.81	2666.01	2272.20	1.2387	7.4275
96	0.87771	0.001040	1.914	402.23	2669.22	2266.98	1.2616	7.4027
98	0.94390	0.001042	1.788	410.66	2672.40	2261.74	1.2844	7.3782
100	1.0142	0.001043	1.672	419.10	2675.57	2256.47	1.3070	7.3541
105	1.2090	0.001047	1.418	440.21	2683.39	2243.18	1.3632	7.2951
110	1.4338	0.001052	1.209	461.36	2691.07	2229.70	1.4187	7.2380
115	1.6918	0.001056	1.036	482.55	2698.58	2216.03	1.4735	7.1827
120	1.9867	0.001060	0.8913	503.78	2705.93	2202.15	1.5278	7.1291
125	2.3222	0.001065	0.7701	525.06	2713.11	2188.04	1.5815	7.0770
130	2.7026	0.001070	0.6681	546.39	2720.09	2173.70	1.6346	7.0264
135	3.1320	0.001075	0.5818	567.77	2726.87	2159.10	1.6872	6.9772
140	3.6150	0.001080	0.5085	589.20	2733.44	2144.24	1.7393	6.9293
145	4.1564	0.001085	0.4460	610.69	2739.80	2129.10	1.7909	6.8826
150	4.7610	0.001091	0.3925	632.25	2745.92	2113.67	1.8420	6.8370
155	5.4342	0.001096	0.3465	653.88	2751.80	2097.92	1.8926	6.7926
160	6.1814	0.001102	0.3068	675.57	2757.43	2081.86	1.9428	6.7491
165	7.0082	0.001108	0.2725	697.35	2762.80	2065.45	1.9926	6.7066
170	7.9205	0.001114	0.2426	719.21	2767.89	2048.69	2.0419	6.6649
175	8.9245	0.001121	0.2166	741.15	2772.70	2031.55	2.0909	6.6241
180	10.026	0.001127	0.1939	763.19	2777.22	2014.03	2.1395	6.5841
185	11.233	0.001134	0.1739	785.32	2781.43	1996.10	2.1878	6.5447
190	12.550	0.001141	0.1564	807.57	2785.31	1977.75	2.2358	6.5060
195	13.986	0.001149	0.1409	829.92	2788.86	1958.94	2.2834	6.4679
200	15.547	0.001157	0.1272	852.39	2792.06	1939.67	2.3308	6.4303
205	17.240	0.001164	0.1151	874.99	2794.90	1919.90	2.3779	6.3932
210	19.074	0.001173	0.1043	897.73	2797.35	1899.62	2.4248	6.3565
215	21.056	0.001181	0.09469	920.61	2799.41	1878.80	2.4714	6.3202
220	23.193	0.001190	0.08610	943.64	2801.05	1857.41	2.5178	6.2842
225	25.494	0.001199	0.07841	966.84	2802.26	1835.42	2.5641	6.2485
230	27.968	0.001209	0.07151	990.21	2803.01	1812.80	2.6102	6.2131
235	30.622	0.001219	0.06530	1013.77	2803.28	1789.52	2.6561	6.1777
240	33.467	0.001229	0.05971	1037.52	2803.06	1765.54	2.7019	6.1425
245	36.509	0.001240	0.05466	1061.49	2802.31	1740.82	2.7477	6.1074
250	39.759	0.001252	0.05009	1085.69	2801.01	1715.33	2.7934	6.0722
255	43.227	0.001264	0.04594	1110.13	2799.13	1689.01	2.8391	6.0370
260	46.921	0.001276	0.04218	1134.83	2796.64	1661.82	2.8847	6.0017
265	50.851	0.001289	0.03875	1159.81	2793.51	1633.70	2.9304	5.9662
270	55.028	0.001303	0.03562	1185.09	2789.69	1604.60	2.9762	5.9304
275	59.463	0.001318	0.03277	1210.70	2785.14	1574.44	3.0221	5.8943
280	64.165	0.001333	0.03015	1236.67	2779.82	1543.15	3.0681	5.8578
285	69.145	0.001349	0.02776	1263.02	2773.67	1510.65	3.1143	5.8208

Table 4.6 (cont.)

t (°C)	p (bar)	v' (m^3/kg)	v'' (m^3/kg)	h' (kJ/kg)	h'' (kJ/kg)	Δh_v (kJ/kg)	s' (kJ/(kgK))	s'' (kJ/(kgK))
290	74.416	0.001366	0.02556	1289.80	2766.63	1476.84	3.1608	5.7832
295	79.990	0.001385	0.02353	1317.03	2758.63	1441.60	3.2076	5.7449
300	85.877	0.001404	0.02166	1344.77	2749.57	1404.80	3.2547	5.7058
305	92.092	0.001425	0.01994	1373.07	2739.38	1366.30	3.3024	5.6656
310	98.647	0.001448	0.01834	1402.00	2727.92	1325.92	3.3506	5.6243
315	105.56	0.001472	0.01686	1431.63	2715.08	1283.45	3.3994	5.5816
320	112.84	0.001499	0.01548	1462.05	2700.67	1238.62	3.4491	5.5373
325	120.51	0.001528	0.01419	1493.37	2684.48	1191.11	3.4997	5.4911
330	128.58	0.001561	0.01298	1525.74	2666.25	1140.51	3.5516	5.4425
335	137.07	0.001597	0.01185	1559.34	2645.60	1086.26	3.6048	5.3910
340	146.00	0.001638	0.01078	1594.45	2622.07	1027.62	3.6599	5.3359
345	155.40	0.001685	0.009770	1631.44	2595.01	963.57	3.7175	5.2763
350	165.29	0.001740	0.008801	1670.86	2563.59	892.73	3.7783	5.2109
355	175.70	0.001808	0.007866	1713.71	2526.45	812.74	3.8438	5.1377
360	186.66	0.001895	0.006945	1761.49	2480.99	719.50	3.9164	5.0527
365	198.22	0.002016	0.006004	1817.59	2422.00	604.41	4.0010	4.9482
370	210.43	0.002222	0.004946	1892.64	2333.50	440.86	4.1142	4.7996
373.946	220.64	0.003106	0.003106	2087.55	2087.55	0.00	4.4120	4.4120

for triatomic gases $\varkappa = 1.30$. The average specific heat is the integral mean value defined by

$$[c_p]_{t_1}^{t_2} = \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} c_p \, dt \,,$$

$$[c_v]_{t_1}^{t_2} = \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} c_v \, dt \,. \qquad (4.74)$$

From (4.71) and (4.72) it follows for the change of internal energy and enthalpy that

$$u_2 - u_1 = [c_v]_{t_1}^{t_2}(t_2 - t_1) = [c_v]_0^{t_2} t_2 - [c_v]_0^{t_1} t_1 \qquad (4.75)$$

and

$$h_2 - h_1 = [c_p]_{t_1}^{t_2}(t_2 - t_1) = [c_p]_0^{t_2} t_2 - [c_p]_0^{t_1} t_1 \,. \qquad (4.76)$$

Numerical values for $[c_p]_0^t$ and $[c_v]_0^t$ can be determined from the average molar specific heats given in Table 4.8.

Taking into account (4.71) and (4.60), the specific entropy arises from (4.29) as

$$ds = \frac{du + p\,dv}{T} = c_v \frac{dT}{T} + R \frac{dv}{v} \,, \qquad (4.77)$$

or after integration with $c_v = \text{const.}$ as

$$s_2 - s_1 = c_v \ln \frac{T_2}{T_1} + R \ln \frac{v_2}{v_1} \,. \qquad (4.78)$$

The integration of (4.32) with constant c_p leads to the equivalent equation

$$s_2 - s_1 = c_p \ln \frac{T_2}{T_1} + R \ln \frac{p_2}{p_1} \,. \qquad (4.79)$$

Real Gases and Vapors

The caloric properties of real gases and vapors are usually determined by measurements, but it is also possible to derive them, apart from an initial value, from equation of states. They are displayed in tables and diagrams as $u = u(v, T)$, $h = h(p, T)$, $s = s(p, T)$, $c_v = c_v(v, T)$,

Table 4.7 Specific heats of air at different pressures calculated with the equation of state (after [4.11])

p (bar)		1	25	50	100	150	200	300	
$t = 0\,°C$	$c_p =$	1.0065	1.0579	1.1116	1.2156	1.3022	1.3612	1.4087	kJ/(kgK)
$t = 50\,°C$	$c_p =$	1.0080	1.0395	1.0720	1.1335	1.1866	1.2288	1.2816	kJ/(kgK)
$t = 100\,°C$	$c_p =$	1.0117	1.0330	1.0549	1.0959	1.1316	1.1614	1.2045	kJ/(kgK)

Table 4.8 Mean molar specific heats $[\bar{C}_p]_0^t$ of ideal gases in kJ/(kmolK) between 0 °C and t °C. The mean molar specific heat $[\bar{C}_v]_0^t$ is determined by subtracting the value of the universal gas constant 8.3143 kJ/(kmolK) from the numerical values given in the table. For the conversion to 1 kg the numerical values have to be divided by the molar weights given in the last line

t (°C)	$[\bar{C}_p]_0^t$ (kJ/(kmolK))							
	H_2	N_2	O_2	CO	H_2O	CO_2	Air	NH_3
0	28.6202	29.0899	29.2642	29.1063	33.4708	35.9176	29.0825	34.99
100	28.9427	29.1151	29.5266	29.1595	33.7121	38.1699	29.1547	36.37
200	29.0717	29.1992	29.9232	29.2882	34.0831	40.1275	29.3033	38.13
300	29.1362	29.3504	30.3871	29.4982	34.5388	41.8299	29.5207	40.02
400	29.1886	29.5632	30.8669	29.7697	35.0485	43.3299	29.7914	41.98
500	29.2470	29.8209	31.3244	30.0805	35.5888	44.6584	30.0927	44.04
600	29.3176	30.1066	31.7499	30.4080	36.1544	45.8462	30.4065	46.09
700	29.4083	30.4006	32.1401	30.7356	36.7415	46.9063	30.7203	48.01
800	29.5171	30.6947	32.4920	31.0519	37.3413	47.8609	31.0265	49.85
900	29.6461	30.9804	32.8151	31.3571	37.9482	48.7231	31.3205	51.53
1000	29.7892	31.2548	33.1094	31.6454	38.5570	49.5017	31.5999	53.08
1100	29.9485	31.5181	33.3781	31.9198	39.1621	50.2055	31.8638	54.50
1200	30.1158	31.7673	33.6245	32.1717	39.7583	50.8522	32.1123	55.84
1300	30.2891	31.9998	33.8548	32.4097	40.3418	51.4373	32.3458	57.06
1400	30.4705	32.2182	34.0723	32.6308	40.9127	51.9783	32.5651	58.14
1500	30.6540	32.4255	34.2771	32.8380	41.4675	52.4710	32.7713	59.19
1600	30.8394	32.6187	34.4690	33.0312	42.0042	52.9285	32.9653	60.20
1700	31.0248	32.7979	34.6513	33.2103	42.5229	53.3508	33.1482	61.12
1800	31.2103	32.9688	34.8305	33.3811	43.0254	53.7423	33.3209	61.95
1900	31.3937	33.1284	35.0000	33.5379	43.5081	54.1030	33.4843	62.75
2000	31.5751	33.2797	35.1664	33.6890	43.9745	54.4418	33.6392	63.46
M (kg/kmol)	2.01588	28.01340	31.999	28.01040	18.01528	44.00980	28.953	17.03052

and $c_p = c_p(p, T)$. Often computer software is necessary to analyze equations of state.

For vapors it holds that the enthalpy h'' of the saturated vapor differs from the enthalpy h' of the saturated boiling liquid at p, $T = \text{const.}$ by the enthalpy of vaporization

$$\Delta h_v = h'' - h' \,, \tag{4.80}$$

which decreases with increasing temperature and reaches zero at the critical point, where $h'' = h'$. The enthalpy of wet vapor is

$$h = (1-x)h' + xh'' = h' + x\Delta h_v \,. \tag{4.81}$$

Accordingly, the internal energy is

$$u = (1-x)u' + xu'' = u' + x(u'' - u') \tag{4.82}$$

and the entropy

$$s = (1-x)s' + xs'' = s' + x\Delta h_v/T \,, \tag{4.83}$$

because enthalpy of vaporization and entropy of vaporization $s'' - s'$ are related through

$$\Delta h_v = T(s'' - s') \,. \tag{4.84}$$

According to the Clausius–Clapeyron relation, the enthalpy of vaporization with gradient $\mathrm{d}p/\mathrm{d}T$ is connected to the liquid–vapor saturation curve $p(T)$ via

$$\Delta h_v = T(v'' - v')\frac{\mathrm{d}p}{\mathrm{d}T} \,, \tag{4.85}$$

with T being the saturation temperature at pressure p. This relationship can be used to calculate the remaining quantity when two out of the three quantities Δh_v, $v'' - v'$, and $\mathrm{d}p/\mathrm{d}T$ are known.

If properties do not have to be calculated continuously or if powerful computers are not available,

Table 4.9 Properties of water and superheated water vapor (after [4.10])

$p \rightarrow$	1 bar t_s = 99.61 °C			5 bar t_s = 151.884 °C			10 bar t_s = 179.89 °C			15 bar t_s = 198.330 °C		
	v''	h''	s''	v''	h''	s''	v''	h''	s''	v''	h''	s''
	1.69402	**2674.9**	**7.3588**	**0.37480**	**2748.1**	**6.8206**	**0.19435**	**2777.1**	**6.5850**	**0.13170**	**2791.0**	**6.4431**
t (°C)	v (m^3/kg)	h (kJ/kg)	s (kJ/(kgK))	v (m^3/kg)	h (kJ/kg)	s (kJ/(kgK))	v (m^3/kg)	h (kJ/kg)	s (kJ/(kgK))	v (m^3/kg)	h (kJ/kg)	s (kJ/(kgK))
0	0.001000	0.06	−0.0001	0.001000	0.47	−0.0001	0.001000	0.98	−0.0001	0.000999	1.48	−0.0001
10	0.001000	42.12	0.1511	0.001000	42.51	0.1510	0.001000	42.99	0.1510	0.001000	43.48	0.1510
20	0.001002	84.01	0.2965	0.001002	84.39	0.2964	0.001001	84.86	0.2963	0.001001	85.33	0.2962
40	0.001008	167.62	0.5724	0.001008	167.98	0.5722	0.001007	168.42	0.5720	0.001007	168.86	0.5719
60	0.001017	251.22	0.8312	0.001017	251.56	0.8310	0.001017	251.98	0.8307	0.001016	252.40	0.8304
80	0.001029	334.99	1.0754	0.001029	335.31	1.0751	0.001029	335.71	1.0748	0.001028	336.10	1.0744
100	1.695959	2675.77	7.3610	0.001043	419.40	1.3067	0.001043	419.77	1.3063	0.001043	420.15	1.3059
120	1.793238	2716.61	7.4676	0.001060	504.00	1.5275	0.001060	504.35	1.5271	0.001060	504.70	1.5266
140	1.889133	2756.70	7.5671	0.001080	589.29	1.7391	0.001079	589.61	1.7386	0.001079	589.94	1.7381
160	1.984139	2796.42	7.6610	0.383660	2767.38	6.8655	0.001102	675.80	1.9423	0.001101	676.09	1.9417
180	2.078533	2835.97	7.7503	0.404655	2812.45	6.9672	0.194418	2777.43	6.5857	0.001127	763.44	2.1389
200	2.172495	2875.48	7.8356	0.425034	2855.90	7.0611	0.206004	2828.27	6.6955	0.132441	2796.02	6.4537
220	2.266142	2915.02	7.9174	0.445001	2898.40	7.1491	0.216966	2875.55	6.7934	0.140630	2850.19	6.5658
240	2.359555	2954.66	7.9962	0.464676	2940.31	7.2324	0.227551	2920.98	6.8837	0.148295	2900.00	6.6649
260	2.452789	2994.45	8.0723	0.484135	2981.88	7.3119	0.237871	2965.23	6.9683	0.155637	2947.45	6.7556
280	2.545883	3034.40	8.1458	0.503432	3023.28	7.3881	0.247998	3008.71	7.0484	0.162752	2993.37	6.8402
300	2.638868	3074.54	8.2171	0.522603	3064.60	7.4614	0.257979	3051.70	7.1247	0.169699	3038.27	6.9199
320	2.731763	3114.89	8.2863	0.541675	3105.93	7.5323	0.267848	3094.40	7.1979	0.176521	3082.48	6.9957
340	2.824585	3155.45	8.3536	0.560667	3147.32	7.6010	0.277629	3136.93	7.2685	0.183245	3126.25	7.0683
360	2.917346	3196.24	8.4190	0.579594	3188.83	7.6676	0.287339	3179.39	7.3366	0.189893	3169.75	7.1381
380	3.010056	3237.27	8.4828	0.598467	3230.48	7.7323	0.296991	3221.86	7.4026	0.196478	3213.09	7.2055
400	3.102722	3278.54	8.5451	0.617294	3272.29	7.7954	0.306595	3264.39	7.4668	0.203012	3256.37	7.2708
420	3.195351	3320.06	8.6059	0.636083	3314.29	7.8569	0.316158	3307.01	7.5292	0.209504	3299.64	7.3341
440	3.287948	3361.83	8.6653	0.654838	3356.49	7.9169	0.325687	3349.76	7.5900	0.215960	3342.96	7.3957
460	3.380516	3403.86	8.7234	0.673565	3398.90	7.9756	0.335186	3392.66	7.6493	0.222385	3386.37	7.4558
480	3.473061	3446.15	8.7803	0.692267	3441.54	8.0329	0.344659	3435.74	7.7073	0.228784	3429.90	7.5143
500	3.565583	3488.71	8.8361	0.710947	3484.41	8.0891	0.354110	3479.00	7.7640	0.235160	3473.57	7.5716
520	3.658087	3531.53	8.8907	0.729607	3527.52	8.1442	0.363541	3522.47	7.8195	0.241515	3517.40	7.6275
540	3.750573	3574.63	8.9444	0.748250	3570.87	8.1981	0.372955	3566.15	7.8739	0.247854	3561.41	7.6823
560	3.843045	3618.00	8.9971	0.766878	3614.48	8.2511	0.382354	3610.05	7.9272	0.254176	3605.61	7.7360
580	3.935503	3661.65	9.0489	0.785493	3658.34	8.3031	0.391738	3654.19	7.9795	0.260485	3650.02	7.7887

Table 4.9 (cont.)

$p \rightarrow$	1 bar t_s = 99.61 °C			5 bar t_s = 151.884 °C			10 bar t_s = 179.89 °C			15 bar t_s = 198.330 °C		
	v''	h''	s''	v''	h''	s''	v''	h''	s''	v''	h''	s''
	1.69402	**2674.9**	**7.3588**	**0.37480**	**2748.1**	**6.8206**	**0.19435**	**2777.1**	**6.5850**	**0.13170**	**2791.0**	**6.4431**
t (°C)	v (m³/kg)	h (kJ/kg)	s (kJ/(kgK))	v (m³/kg)	h (kJ/kg)	s (kJ/(kgK))	v (m³/kg)	h (kJ/kg)	s (kJ/(kgK))	v (m³/kg)	h (kJ/kg)	s (kJ/(kgK))
600	4.027949	3705.57	9.0998	0.804095	3702.46	8.3543	0.401111	3698.56	8.0309	0.266781	3694.64	7.8404
620	4.120384	3749.77	9.1498	0.822687	3746.84	8.4045	0.410472	3743.17	8.0815	0.273066	3739.48	7.8912
640	4.212810	3794.26	9.1991	0.841269	3791.49	8.4539	0.419824	3788.03	8.1311	0.279341	3784.55	7.9411
660	4.305227	3839.02	9.2476	0.859842	3836.41	8.5026	0.429167	3833.14	8.1800	0.285608	3829.86	7.9902
680	4.397636	3884.06	9.2953	0.878406	3881.59	8.5505	0.438502	3878.50	8.2281	0.291866	3875.40	8.0384
700	4.490037	3929.38	9.3424	0.896964	3927.05	8.5977	0.447829	3924.12	8.2755	0.298117	3921.18	8.0860
720	4.582433	3974.99	9.3888	0.915516	3972.77	8.6442	0.457150	3970.00	8.3221	0.304361	3967.22	8.1328
740	4.674822	4020.87	9.4345	0.934061	4018.77	8.6901	0.466465	4016.14	8.3681	0.310600	4013.50	8.1789
760	4.767206	4067.04	9.4796	0.952601	4065.04	8.7353	0.475775	4062.54	8.4135	0.316833	4060.03	8.2244
780	4.859585	4113.48	9.5241	0.971136	4111.58	8.7799	0.485080	4109.21	8.4582	0.323061	4106.82	8.2693
800	4.951960	4160.21	9.5681	0.989667	4158.40	8.8240	0.494380	4156.14	8.5024	0.329284	4153.87	8.3135

Table 4.9 (cont.)

$p \rightarrow$	20 bar t_s = 212.38 °C			25 bar t_s = 223.96 °C			50 bar t_s = 263.94 °C			100 bar t_s = 311.0 °C		
	v''	h''	s''	v''	h''	s''	v''	h''	s''	v''	h''	s''
	0.09958	**2798.4**	**6.3392**	**0.07995**	**2802.0**	**6.2560**	**0.03945**	**2794.2**	**5.9737**	**0.01803**	**2725.5**	**5.6159**
t (°C)	v (m³/kg)	h (kJ/kg)	s (kJ/(kgK))	v (m³/kg)	h (kJ/kg)	s (kJ/(kgK))	v (m³/kg)	h (kJ/kg)	s (kJ/(kgK))	v (m³/kg)	h (kJ/kg)	s (kJ/(kgK))
0	0.000999	1.99	0.0000	0.000999	2.50	0.0000	0.000998	5.03	0.0001	0.000995	10.07	0.0003
10	0.000999	43.97	0.1509	0.000999	44.45	0.1509	0.000998	46.88	0.1506	0.000996	51.72	0.1501
20	0.001001	85.80	0.2961	0.001001	86.27	0.2960	0.001000	88.61	0.2955	0.000997	93.29	0.2944
40	0.001007	169.31	0.5717	0.001007	169.75	0.5715	0.001006	171.96	0.5705	0.001004	176.37	0.5685
60	0.001016	252.82	0.8302	0.001016	253.24	0.8299	0.001015	255.33	0.8286	0.001013	259.53	0.8259
80	0.001028	336.50	1.0741	0.001028	336.90	1.0738	0.001027	338.89	1.0721	0.001024	342.87	1.0689
100	0.001042	420.53	1.3055	0.001042	420.90	1.3051	0.001041	422.78	1.3032	0.001039	426.55	1.2994
120	0.001059	505.05	1.5262	0.001059	505.40	1.5257	0.001058	507.17	1.5235	0.001055	510.70	1.5190
140	0.001079	590.26	1.7376	0.001078	590.59	1.7371	0.001077	592.22	1.7345	0.001074	595.49	1.7294
160	0.001101	676.38	1.9411	0.001101	676.67	1.9405	0.001099	678.14	1.9376	0.001095	681.11	1.9318
180	0.001127	763.69	2.1382	0.001126	763.94	2.1375	0.001124	765.22	2.1341	0.001120	767.81	2.1274
200	0.001156	852.57	2.3301	0.001156	852.77	2.3293	0.001153	853.80	2.3254	0.001148	855.92	2.3177

Table 4.9 (cont.)

$p \rightarrow$	20 bar t_s = 212.38 °C			25 bar t_s = 223.96 °C			50 bar t_s = 263.94 °C			100 bar t_s = 311.0 °C		
	v''	h''	s''	v''	h''	s''	v''	h''	s''	v''	h''	s''
	0.09958	**2798.4**	**6.3392**	**0.07995**	**2802.0**	**6.2560**	**0.03945**	**2794.2**	**5.9737**	**0.01803**	**2725.5**	**5.6159**
t (°C)	v (m^3/kg)	h (kJ/kg)	s (kJ/(kgK))	v (m^3/kg)	h (kJ/kg)	s (kJ/(kgK))	v (m^3/kg)	h (kJ/kg)	s (kJ/(kgK))	v (m^3/kg)	h (kJ/kg)	s (kJ/(kgK))
220	0.102167	2821.67	6.3868	0.001190	943.69	2.5175	0.001187	944.38	2.5129	0.001181	945.87	2.5039
240	0.108488	2877.21	6.4972	0.084437	2852.28	6.3555	0.001227	1037.68	2.6983	0.001219	1038.30	2.6876
260	0.114400	2928.47	6.5952	0.089553	2908.19	6.4624	0.001275	1134.77	2.8839	0.001265	1134.13	2.8708
280	0.120046	2977.21	6.6850	0.094351	2960.16	6.5581	0.042275	2858.08	6.0909	0.001323	1234.82	3.0561
300	0.125501	3024.25	6.7685	0.098932	3009.63	6.6460	0.045347	2925.64	6.2109	0.001398	1343.10	3.2484
320	0.130816	3070.16	6.8472	0.103357	3057.40	6.7279	0.048130	2986.18	6.3148	0.019272	2782.66	5.7131
340	0.136023	3115.28	6.9221	0.107664	3104.01	6.8052	0.050726	3042.36	6.4080	0.021490	2882.06	5.8780
360	0.141147	3159.89	6.9937	0.111881	3149.81	6.8787	0.053188	3095.62	6.4934	0.023327	2962.61	6.0073
380	0.146205	3204.16	7.0625	0.116026	3195.07	6.9491	0.055552	3146.83	6.5731	0.024952	3033.11	6.1170
400	0.151208	3248.23	7.1290	0.120115	3239.96	7.0168	0.057840	3196.59	6.6481	0.026439	3097.38	6.2139
420	0.156167	3292.18	7.1933	0.124156	3284.63	7.0822	0.060068	3245.31	6.7194	0.027829	3157.45	6.3019
440	0.161088	3336.09	7.2558	0.128159	3329.15	7.1455	0.062249	3293.27	6.7877	0.029148	3214.57	6.3831
460	0.165978	3380.02	7.3165	0.132129	3373.62	7.2070	0.064391	3340.68	6.8532	0.030410	3269.53	6.4591
480	0.170841	3424.01	7.3757	0.136072	3418.08	7.2668	0.066501	3387.71	6.9165	0.031629	3322.89	6.5310
500	0.175680	3468.09	7.4335	0.139990	3462.59	7.3251	0.068583	3434.48	6.9778	0.032813	3375.06	6.5993
520	0.180499	3512.30	7.4899	0.143887	3507.17	7.3821	0.070642	3481.06	7.0373	0.033968	3426.31	6.6648
540	0.185300	3556.64	7.5451	0.147766	3551.85	7.4377	0.072681	3527.54	7.0952	0.035098	3476.87	6.7277
560	0.190085	3601.15	7.5992	0.151629	3596.67	7.4922	0.074703	3573.96	7.1516	0.036208	3526.90	6.7885
580	0.194856	3645.84	7.6522	0.155477	3641.64	7.5455	0.076710	3620.38	7.2066	0.037300	3576.52	6.8474
600	0.199614	3690.71	7.7042	0.159313	3686.76	7.5978	0.078703	3666.83	7.2604	0.038377	3625.84	6.9045
620	0.204362	3735.78	7.7552	0.163138	3732.07	7.6491	0.080684	3713.34	7.3131	0.039442	3674.95	6.9601
640	0.209099	3781.07	7.8054	0.166953	3777.57	7.6995	0.082655	3759.94	7.3647	0.040494	3723.89	7.0143
660	0.213827	3826.57	7.8547	0.170758	3823.27	7.7490	0.084616	3806.65	7.4153	0.041536	3772.73	7.0672
680	0.218547	3872.29	7.9032	0.174556	3869.17	7.7976	0.086569	3853.48	7.4650	0.042569	3821.51	7.1189
700	0.223260	3918.24	7.9509	0.178346	3915.30	7.8455	0.088515	3900.45	7.5137	0.043594	3870.27	7.1696
720	0.227966	3964.43	7.9978	0.182129	3961.64	7.8927	0.090453	3947.58	7.5617	0.044612	3919.04	7.2192
740	0.232667	4010.86	8.0441	0.185907	4008.21	7.9391	0.092385	3994.88	7.6088	0.045623	3967.85	7.2678
760	0.237361	4057.52	8.0897	0.189679	4055.01	7.9848	0.094312	4042.35	7.6552	0.046629	4016.72	7.3156
780	0.242051	4104.43	8.1347	0.193446	4102.04	8.0299	0.096234	4090.02	7.7009	0.047629	4065.68	7.3625
800	0.246737	4151.59	8.1791	0.197208	4149.32	8.0744	0.098151	4137.87	7.7459	0.048624	4114.73	7.4087

Table 4.9 (cont.)

$p \rightarrow$	150 bar t_s = 342.16 °C			200 bar t_s = 365.765 °C			250 bar			300 bar		
	v''	h''	s''	v''	h''	s''						
t	0.01034	2610.9	5.3108	0.00586	2411.4	4.9299						
(°C)	v (m³/kg)	h (kJ/kg)	s (kJ/(kgK))	v (m³/kg)	h (kJ/kg)	s (kJ/(kgK))	v (m³/kg)	h (kJ/kg)	s (kJ/(kgK))	v (m³/kg)	h (kJ/kg)	s (kJ/(kgK))
0	0.000993	15.07	0.0004	0.000990	20.03	0.0005	0.000988	24.96	0.0004	0.000986	29.86	0.0003
10	0.000993	56.52	0.1495	0.000991	61.30	0.1489	0.000989	66.06	0.1482	0.000987	70.79	0.1474
20	0.000995	97.94	0.2932	0.000993	102.57	0.2921	0.000991	107.18	0.2909	0.000989	111.78	0.2897
40	0.001001	180.78	0.5666	0.000999	185.17	0.5646	0.000997	189.54	0.5627	0.000995	193.91	0.5607
60	0.001011	263.71	0.8233	0.001008	267.89	0.8207	0.001006	272.07	0.8181	0.001004	276.24	0.8156
80	0.001022	346.85	1.0657	0.001020	350.83	1.0625	0.001018	354.82	1.0593	0.001016	358.80	1.0562
100	0.001036	430.32	1.2956	0.001034	434.10	1.2918	0.001031	437.88	1.2881	0.001029	441.67	1.2845
120	0.001052	514.25	1.5147	0.001050	517.81	1.5104	0.001047	521.38	1.5061	0.001045	524.97	1.5019
140	0.001071	598.79	1.7244	0.001068	602.11	1.7195	0.001065	605.45	1.7147	0.001062	608.80	1.7099
160	0.001092	684.12	1.9261	0.001089	687.15	1.9205	0.001085	690.22	1.9150	0.001082	693.31	1.9097
180	0.001116	770.46	2.1209	0.001112	773.16	2.1146	0.001108	775.90	2.1084	0.001105	778.68	2.1023
200	0.001144	858.12	2.3102	0.001139	860.39	2.3030	0.001135	862.73	2.2959	0.001130	865.14	2.2890
220	0.001175	947.49	2.4952	0.001170	949.22	2.4868	0.001164	951.06	2.4787	0.001159	952.99	2.4709
240	0.001212	1039.13	2.6774	0.001205	1040.14	2.6675	0.001199	1041.31	2.6581	0.001193	1042.62	2.6490
260	0.001256	1133.83	2.8584	0.001247	1133.83	2.8466	0.001239	1134.08	2.8355	0.001231	1134.57	2.8248
280	0.001310	1232.79	3.0406	0.001298	1231.29	3.0261	0.001287	1230.24	3.0125	0.001277	1229.56	2.9997
300	0.001378	1338.06	3.2275	0.001361	1334.14	3.2087	0.001346	1331.06	3.1915	0.001332	1328.66	3.1756
320	0.001473	1453.85	3.4260	0.001445	1445.30	3.3993	0.001421	1438.72	3.3761	0.001401	1433.51	3.3554
340	0.001631	1592.27	3.6553	0.001569	1571.52	3.6085	0.001526	1557.48	3.5729	0.001493	1547.07	3.5437
360	0.012582	2769.56	5.5654	0.001825	1740.13	3.8787	0.001697	1698.63	3.7993	0.001628	1675.57	3.7498
380	0.014289	2884.61	5.7445	0.008258	2659.19	5.3144	0.002218	1935.67	4.1670	0.001873	1838.26	4.0026
400	0.015671	2975.55	5.8817	0.009950	2816.84	5.5525	0.006005	2578.59	5.1399	0.002796	2152.37	4.4750
420	0.016875	3053.94	5.9965	0.011199	2928.51	5.7160	0.007579	2769.45	5.4196	0.004921	2552.87	5.0625
440	0.017965	3124.58	6.0970	0.012246	3020.26	5.8466	0.008697	2897.06	5.6013	0.006228	2748.86	5.3416
460	0.018974	3190.02	6.1875	0.013170	3100.57	5.9577	0.009617	2999.20	5.7426	0.007193	2883.84	5.5284
480	0.019924	3251.76	6.2706	0.014011	3173.45	6.0558	0.010418	3087.11	5.8609	0.007992	2991.99	5.6740
500	0.020828	3310.79	6.3479	0.014793	3241.19	6.1445	0.011142	3165.92	5.9642	0.008690	3084.79	5.7956
520	0.021696	3367.79	6.4207	0.015530	3305.21	6.2263	0.011810	3238.48	6.0569	0.009320	3167.67	5.9015

Table 4.9 (cont.)

$p \rightarrow$	150 bar t_s = 342.16 °C				200 bar t_s = 365.765 °C				250 bar				300 bar			
	v''	h''	s''		v''	h''	s''									
	0.01034	2610.9	5.3108		0.00586	2411.4	4.9299									
t	v	h	s		v	h	s		v	h	s		v	h	s	
(°C)	(m^3/kg)	(kJ/kg)	(kJ/(kgK))		(m^3/kg)	(kJ/kg)	(kJ/(kgK))		(m^3/kg)	(kJ/kg)	(kJ/(kgK))		(m^3/kg)	(kJ/kg)	(kJ/(kgK))	
540	0.022535	3423.22	6.4897		0.016231	3366.45	6.3026		0.012435	3306.55	6.1416		0.009899	3243.71	5.9962	
560	0.023350	3477.46	6.5556		0.016904	3425.57	6.3744		0.013028	3371.29	6.2203		0.010442	3314.82	6.0826	
580	0.024144	3530.75	6.6188		0.017554	3483.05	6.4426		0.013595	3433.49	6.2941		0.010955	3382.25	6.1626	
600	0.024921	3583.31	6.6797		0.018184	3539.23	6.5077		0.014140	3493.69	6.3638		0.011444	3446.87	6.2374	
620	0.025683	3635.28	6.7386		0.018799	3594.37	6.5701		0.014667	3552.32	6.4302		0.011914	3509.28	6.3081	
640	0.026432	3686.79	6.7956		0.019399	3648.69	6.6303		0.015179	3609.69	6.4937		0.012368	3569.91	6.3752	
660	0.027171	3737.95	6.8510		0.019987	3702.35	6.6884		0.015678	3666.03	6.5548		0.012808	3629.12	6.4394	
680	0.027899	3788.82	6.9050		0.020564	3755.46	6.7447		0.016165	3721.54	6.6136		0.013236	3687.16	6.5009	
700	0.028619	3839.48	6.9576		0.021133	3808.15	6.7994		0.016643	3776.37	6.6706		0.013654	3744.24	6.5602	
720	0.029332	3889.99	7.0090		0.021693	3860.50	6.8527		0.017113	3830.64	6.7258		0.014063	3800.53	6.6175	
740	0.030037	3940.39	7.0592		0.022246	3912.57	6.9046		0.017575	3884.47	6.7794		0.014464	3856.17	6.6729	
760	0.030736	3990.72	7.1084		0.022792	3964.43	6.9553		0.018030	3937.92	6.8317		0.014858	3911.27	6.7268	
780	0.031430	4041.03	7.1566		0.023333	4016.13	7.0048		0.018479	3991.08	6.8826		0.015246	3965.93	6.7792	
800	0.032118	4091.33	7.2039		0.023869	4067.73	7.0534		0.018922	4044.00	6.9324		0.015629	4020.23	6.8303	

Table 4.9 (cont.)

$p \rightarrow$	350 bar			400 bar			500 bar		
t (°C)	v (m^3/kg)	h (kJ/kg)	s (kJ/(kgK))	v (m^3/kg)	h (kJ/kg)	s (kJ/(kgK))	v (m^3/kg)	h (kJ/kg)	s (kJ/(kgK))
0	0.000983	34.72	0.0001	0.000981	39.56	−0.0002	0.000977	49.13	−0.0010
10	0.000984	75.49	0.1466	0.000982	80.17	0.1458	0.000978	89.46	0.1440
20	0.000987	116.35	0.2884	0.000985	120.90	0.2872	0.000980	129.96	0.2845
40	0.000993	198.27	0.5588	0.000991	202.61	0.5568	0.000987	211.27	0.5528
60	0.001002	280.40	0.8130	0.001000	284.56	0.8105	0.000996	292.86	0.8054
80	0.001013	362.78	1.0531	0.001011	366.76	1.0501	0.001007	374.71	1.0440
100	0.001027	445.47	1.2809	0.001024	449.26	1.2773	0.001020	456.87	1.2703
120	0.001042	528.56	1.4978	0.001040	532.17	1.4937	0.001035	539.41	1.4858
140	0.001060	612.18	1.7052	0.001057	615.57	1.7006	0.001052	622.40	1.6917
160	0.001079	696.44	1.9044	0.001076	699.59	1.8992	0.001070	705.95	1.8891
180	0.001101	781.51	2.0964	0.001098	784.37	2.0906	0.001091	790.20	2.0793
200	0.001126	867.60	2.2823	0.001122	870.12	2.2758	0.001115	875.31	2.2631
220	0.001155	955.00	2.4632	0.001150	957.10	2.4558	0.001141	961.50	2.4415
240	0.001187	1044.06	2.6402	0.001181	1045.62	2.6317	0.001171	1049.05	2.6155
260	0.001224	1135.25	2.8145	0.001217	1136.11	2.8047	0.001204	1138.29	2.7861
280	0.001268	1229.20	2.9875	0.001259	1229.13	2.9760	0.001243	1229.67	2.9543
300	0.001320	1326.81	3.1608	0.001308	1325.41	3.1469	0.001288	1323.74	3.1214
320	0.001384	1429.36	3.3367	0.001368	1426.02	3.3195	0.001341	1421.22	3.2885
340	0.001466	1538.97	3.5184	0.001443	1532.52	3.4960	0.001405	1523.05	3.4574
360	0.001579	1659.61	3.7119	0.001542	1647.62	3.6807	0.001485	1630.63	3.6300
380	0.001755	1800.51	3.9309	0.001682	1776.72	3.8814	0.001588	1746.51	3.8101
400	0.002106	1988.43	4.2140	0.001911	1931.13	4.1141	0.001731	1874.31	4.0028
420	0.003082	2291.32	4.6570	0.002361	2136.30	4.4142	0.001940	2020.07	4.2161
440	0.004413	2571.64	5.0561	0.003210	2394.03	4.7807	0.002266	2190.53	4.4585
460	0.005436	2753.55	5.3079	0.004149	2613.32	5.0842	0.002745	2380.52	4.7212
480	0.006246	2888.06	5.4890	0.004950	2777.18	5.3048	0.003319	2563.86	4.9680
500	0.006933	2998.02	5.6331	0.005625	2906.69	5.4746	0.003889	2722.52	5.1759
420	0.007540	3093.08	5.7546	0.006213	3015.42	5.6135	0.004417	2857.36	5.3482
540	0.008089	3178.24	5.8606	0.006740	3110.69	5.7322	0.004896	2973.16	5.4924
560	0.008597	3256.46	5.9557	0.007221	3196.67	5.8366	0.005332	3075.37	5.6166
580	0.009073	3329.64	6.0425	0.007669	3276.01	5.9308	0.005734	3167.66	5.7261
600	0.009523	3399.02	6.1229	0.008089	3350.43	6.0170	0.006109	3252.61	5.8245
620	0.009953	3465.45	6.1981	0.008488	3421.10	6.0970	0.006461	3332.05	5.9145
640	0.010365	3529.55	6.2691	0.008869	3488.82	6.1720	0.006796	3407.21	5.9977
660	0.010763	3591.77	6.3365	0.009235	3554.17	6.2428	0.007115	3478.99	6.0755
680	0.011149	3652.46	6.4008	0.009589	3617.59	6.3100	0.007422	3548.00	6.1487
700	0.011524	3711.88	6.4625	0.009931	3679.42	6.3743	0.007718	3614.76	6.2180
720	0.011889	3770.27	6.5219	0.010264	3739.95	6.4358	0.008004	3679.64	6.2840
740	0.012247	3827.78	6.5793	0.010589	3799.38	6.4951	0.008281	3742.97	6.3471
760	0.012598	3884.58	6.6348	0.010906	3857.91	6.5523	0.008552	3804.99	6.4078
780	0.012942	3940.78	6.6887	0.011217	3915.68	6.6077	0.008816	3865.93	6.4662
800	0.013280	3996.48	6.7411	0.011523	3972.81	6.6614	0.009074	3925.96	6.5226

Table 4.10 Properties of ammonia (NH_3) at saturation (after [4.14])

Temperature	Pressure	Specific volume		Enthalpy		Enthalpy	Entropy	
		liquid	vapor	liquid	vapor	vaporization	liquid	vapor
t (°C)	p (bar)	v' (dm^3/kg)	v'' (dm^3/kg)	h' (kJ/kg)	h'' (kJ/kg)	$\Delta h_v = h'' - h'$ (kJ/kg)	s' (kJ/(kgK))	s'' (kJ/(kgK))
−70	0.10941	1.3798	9007.9	−110.81	1355.6	1466.4	−0.30939	6.9088
−60	0.21893	1.4013	4705.7	−68.062	1373.7	1441.8	−0.10405	6.6602
−50	0.40836	1.4243	2627.8	−24.727	1391.2	1415.9	0.09450	6.4396
−40	0.71692	1.4490	1553.3	19.170	1407.8	1388.6	0.28673	6.2425
−30	1.1943	1.4753	963.96	63.603	1423.3	1359.7	0.47303	6.0651
−20	1.9008	1.5035	623.73	108.55	1437.7	1329.1	0.65376	5.9041
−10	2.9071	1.5336	418.30	154.01	1450.7	1296.7	0.82928	5.7569
0	4.2938	1.5660	289.30	200.00	1462.2	1262.2	1.0000	5.6210
10	6.1505	1.6009	205.43	246.57	1472.1	1225.5	1.1664	5.4946
20	8.5748	1.6388	149.20	293.78	1480.2	1186.4	1.3289	5.3759
30	11.672	1.6802	110.46	341.76	1486.2	1144.4	1.4881	5.2631
40	15.554	1.7258	83.101	390.64	1489.9	1099.3	1.6446	5.1549
50	20.340	1.7766	63.350	440.62	1491.1	1050.5	1.7990	5.0497
60	26.156	1.8340	48.797	491.97	1489.3	997.30	1.9523	4.9458
70	33.135	1.9000	37.868	545.04	1483.9	938.90	2.1054	4.8415
80	41.420	1.9776	29.509	600.34	1474.3	873.97	2.2596	4.7344
90	51.167	2.0714	22.997	658.61	1459.2	800.58	2.4168	4.6213
100	62.553	2.1899	17.820	721.00	1436.6	715.63	2.5797	4.4975
110	75.783	2.3496	13.596	789.68	1403.1	613.39	2.7533	4.3543
120	91.125	2.5941	9.9932	869.92	1350.2	480.31	2.9502	4.1719
130	108.98	3.2021	6.3790	992.02	1239.3	247.30	3.2437	3.8571

At the reference state $t = 0\,°C$ saturated liquid possesses the enthalpy $h' = 200.0\,kJ/kg$ and the specific entropy $s' = 1.0\,kJ/(kgK)$

saturated water tables, in which the results of theoretical and experimental investigations are collected, are used for practical calculations. Such tables are collected in Tables 4.6–4.13, for working fluids important in engineering. Diagrams are advantageous to determine reference values and to display changes of state, e.g., a t–s diagram as shown in Fig. 4.9. Most commonly used in practice are Mollier diagrams, which include the enthalpy as one of the coordinates, see Fig. 4.10.

The specific heat $c_p = (\partial h/\partial T)_p$ of vapor depends, as well as on temperature, also considerably on pressure. In the same way, $c_v = (\partial u/\partial T)_v$ depends, besides on temperature, also on the specific volume. Approaching the saturated vapor line, c_p of the superheated vapor increases considerably with decreasing temperature and tends toward infinity at the critical point. While $c_p - c_v$ is a constant for ideal gases, this is not true for vapors.

4.6.3 Incompressible Fluids

An incompressible fluid is a fluid whose specific volume depends neither on temperature nor on pressure, such that the equation of state is given by $v = \text{const}$. As a good approximation, liquids and solids can generally be considered as incompressible. The specific heats c_p and c_v do not differ for incompressible fluids, $c_p = c_v = c$. Thus the caloric equations of state are

$$\mathrm{d}u = c\,\mathrm{d}T \tag{4.86}$$

and

$$\mathrm{d}h = c\,\mathrm{d}T + v\,\mathrm{d}p\,, \tag{4.87}$$

as well as

$$\mathrm{d}s = c\frac{\mathrm{d}T}{T}\,. \tag{4.88}$$

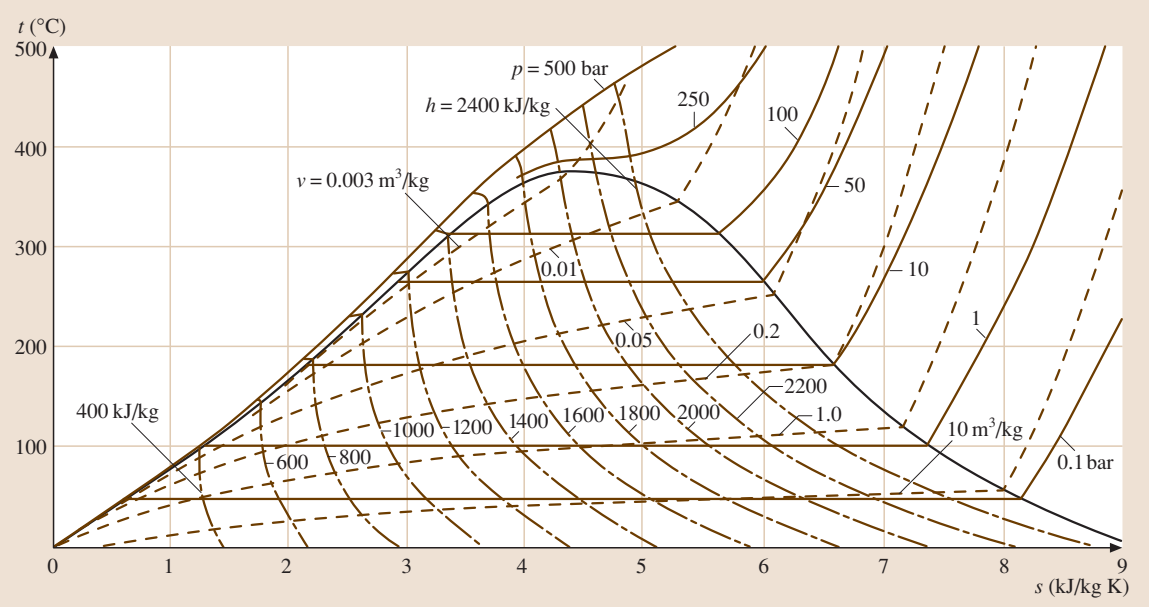


Fig. 4.9 *t–s* diagram of water with curves $p = \text{const}$ (*solid lines*), $v = \text{const}$ (*dashed lines*), and $h = \text{const}$ (*dot* and *dash lines*)

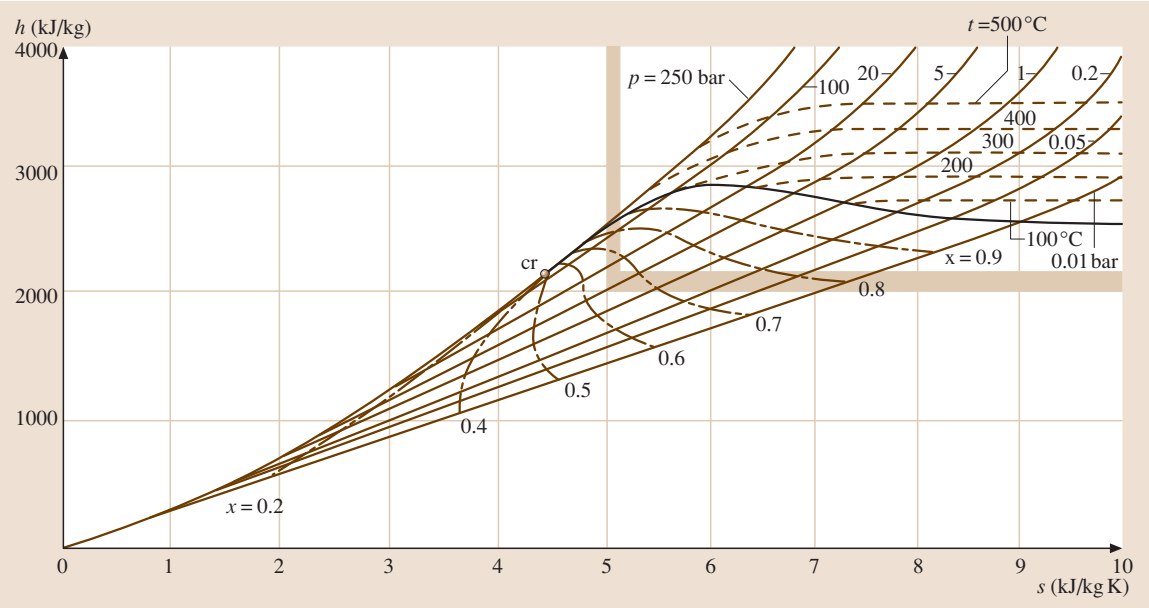


Fig. 4.10 *h–s* diagram of water with curves $p = \text{const}$ (*solid lines*), $t = \text{const}$ (*dashed lines*), and $x = \text{const}$ (*dot* and *dash lines*). The area of interest for the purpose of vapor technology is marked by the *hatched boundary*

Table 4.11 Properties of carbon dioxide (CO_2) at saturation (after [4.15])

Temperature	Pressure	Specific volume		Enthalpy		Enthalpy	Entropy	
		liquid	vapor	liquid	vapor	vaporization	liquid	vapor
t (°C)	p (bar)	v' (dm^3/kg)	v'' (dm^3/kg)	h' (kJ/kg)	h'' (kJ/kg)	$\Delta h_v = h'' - h'$ (kJ/kg)	s' (kJ/(kgK))	s'' (kJ/(kgK))
−55	5.540	0.8526	68.15	83.02	431.0	348.0	0.5349	2.130
−50	6.824	0.8661	55.78	92.93	432.7	339.8	0.5793	2.102
−45	8.319	0.8804	46.04	102.9	434.1	331.2	0.6629	2.075
−40	10.05	0.8957	38.28	112.9	435.3	322.4	0.6658	2.048
−35	12.02	0.9120	32.03	123.1	436.2	313.1	0.7081	2.023
−30	14.28	0.9296	26.95	133.4	436.8	303.4	0.7500	1.998
−25	16.83	0.9486	22.79	143.8	437.0	293.2	0.7915	1.973
−20	19.70	0.9693	19.34	154.5	436.9	282.4	0.8329	1.949
−15	22.91	0.9921	16.47	165.4	436.3	270.9	0.8743	1.924
−10	26.49	1.017	14.05	176.5	435.1	258.6	0.9157	1.898
−5	30.46	1.046	12.00	188.0	433.4	245.3	0.9576	1.872
0	34.85	1.078	10.24	200.0	430.9	230.9	1.000	1.845
5	39.69	1.116	8.724	212.5	427.5	215.0	1.043	1.816
10	45.02	1.161	7.399	225.7	422.9	197.1	1.088	1.785
15	50.87	1.218	6.222	240.0	416.6	176.7	1.136	1.749
20	57.29	1.293	5.150	255.8	407.9	152.0	1.188	1.706
25	64.34	1.408	4.121	274.8	394.5	119.7	1.249	1.650
30	72.14	1.686	2.896	304.6	365.0	60.50	1.343	1.543

Reference points: see footnote of Table 4.10

4.6.4 Solid Materials

Thermal Expansion

Similar to liquids, the influence of pressure on volume in equations of state $V = V(p, T)$ for solids is mostly negligibly small. Nearly all solids expand like liquids with increasing temperature and shrink with decreasing temperature. An exception is water, which has its highest density at 4 °C and expands both at higher and lower temperatures than 4 °C. A Taylor-series expansion with respect to temperature of the equation of state, truncated after the linear term, leads to the volumetric expansion with the cubic volumetric expansion coefficient γ_V (SI unit 1/K)

$$V = V_0 \left[1 + \gamma_V (t - t_0)\right] .$$

Accordingly, the area expansion is

$$A = A_0 \left[1 + \gamma_A (t - t_0)\right]$$

and the length expansion

$$l = l_0 \left[1 + \gamma_L (t - t_0)\right] ,$$

where $\gamma_A = (2/3)\gamma_V$ and $\gamma_L = (1/3)\gamma_V$. Average values for γ_L in the temperature interval between 0 °C and t °C can be derived for some solids from the values in Table 4.14 by dividing the given length change $(l - l_0)/l_0$ by the temperature interval $t - 0\,°C$.

Melting and Sublimation Curve

Within certain limits, each pressure of a liquid corresponds to a temperature at which the liquid is in equilibrium with its solid. This relationship $p(T)$ is determined by the melting curve (Fig. 4.11), whereas the sublimation curve displays the equilibrium between gas and solid. Figure 4.11 includes additionally the liquid–vapor saturation curve. All three curves meet at the triple point at which the solid, the liquid, and the gaseous phase of a substance are in equilibrium with

Table 4.12 Properties of tetrafluoroethane ($C_2H_2F_4$ (R134a)) at saturation (after [4.16, 17])

Temperature	Pressure	Specific volume		Enthalpy		Enthalpy	Entropy	
		liquid	vapor	liquid	vapor	vaporization	liquid	vapor
t (°C)	p (bar)	v' (dm^3/kg)	v'' (dm^3/kg)	h' (kJ/kg)	h'' (kJ/kg)	$\Delta h_v = h'' - h'$ (kJ/kg)	s' (kJ/(kgK))	s'' (kJ/(kgK))
−100	0.0055940	0.63195	25 193	75.362	336.85	261.49	0.43540	1.9456
−95	0.0093899	0.63729	15 435	81.288	339.78	258.50	0.46913	1.9201
−90	0.015241	0.64274	9769.8	87.226	342.76	255.53	0.50201	1.8972
−85	0.023990	0.64831	6370.7	93.182	345.77	252.59	0.53409	1.8766
−80	0.036719	0.65401	4268.2	99.161	348.83	249.67	0.56544	1.8580
−75	0.054777	0.65985	2931.2	105.17	351.91	246.74	0.59613	1.8414
−70	0.079814	0.66583	2059.0	111.20	355.02	243.82	0.62619	1.8264
−65	0.11380	0.67197	1476.5	117.26	358.16	240.89	0.65568	1.8130
−60	0.15906	0.67827	1079.0	123.36	361.31	237.95	0.68462	1.8010
−55	0.21828	0.68475	802.36	129.50	364.48	234.98	0.71305	1.7902
−50	0.29451	0.69142	606.20	135.67	367.65	231.98	0.74101	1.7806
−45	0.39117	0.69828	464.73	141.89	370.83	228.94	0.76852	1.7720
−40	0.51209	0.70537	361.08	148.14	374.00	225.86	0.79561	1.7643
−35	0.66144	0.71268	284.02	154.44	377.17	222.72	0.82230	1.7575
−30	0.84378	0.72025	225.94	160.79	380.32	219.53	0.84863	1.7515
−25	1.0640	0.72809	181.62	167.19	383.45	216.26	0.87460	1.7461
−20	1.3273	0.73623	147.39	173.64	386.55	212.92	0.90025	1.7413
−15	1.6394	0.74469	120.67	180.14	389.63	209.49	0.92559	1.7371
−10	2.0060	0.75351	99.590	186.70	392.66	205.97	0.95065	1.7334
−5	2.4334	0.76271	82.801	193.32	395.66	202.34	0.97544	1.7300
0	2.9280	0.77233	69.309	200.00	398.60	198.60	1.0000	1.7271
5	3.4966	0.78243	58.374	206.75	401.49	194.74	1.0243	1.7245
10	4.1461	0.79305	49.442	213.58	404.32	190.74	1.0485	1.7221
15	4.8837	0.80425	42.090	220.48	407.07	186.59	1.0724	1.7200
20	5.7171	0.81610	35.997	227.47	409.75	182.28	1.0962	1.7180
25	6.6538	0.82870	30.912	234.55	412.33	177.79	1.1199	1.7162
30	7.7020	0.84213	26.642	241.72	414.82	173.10	1.1435	1.7145
35	8.8698	0.85653	23.033	249.01	417.19	168.18	1.1670	1.7128
40	10.166	0.87204	19.966	256.41	419.43	163.02	1.1905	1.7111
45	11.599	0.88885	17.344	263.94	421.52	157.58	1.2139	1.7092
50	13.179	0.90719	15.089	271.62	423.44	151.81	1.2375	1.7072
55	14.915	0.92737	13.140	279.47	425.15	145.68	1.2611	1.7050
60	16.818	0.94979	11.444	287.50	426.63	139.12	1.2848	1.7024
65	18.898	0.97500	9.9604	295.76	427.82	132.06	1.3088	1.6993
70	21.168	1.0038	8.6527	304.28	428.65	124.37	1.3332	1.6956
75	23.641	1.0372	7.4910	313.13	429.03	115.90	1.3580	1.6909
80	26.332	1.0773	6.4483	322.39	428.81	106.42	1.3836	1.6850
85	29.258	1.1272	5.4990	332.22	427.76	95.536	1.4104	1.6771
90	32.442	1.1936	4.6134	342.93	425.42	82.487	1.4390	1.6662
95	35.912	1.2942	3.7434	355.25	420.67	65.423	1.4715	1.6492
100	39.724	1.5357	2.6809	373.30	407.68	34.385	1.5188	1.6109

Reference points: see footnote of Table 4.10

Table 4.13 Properties of chlorodifluoromethane (CHF_3Cl (R22)) at saturation (after [4.18])

Temperature	Pressure	Specific volume		Enthalpy		Enthalpy	Entropy	
		liquid	vapor	liquid	vapor	vaporization	liquid	vapor
t (°C)	p (bar)	v' (dm^3/kg)	v'' (dm^3/kg)	h' (kJ/kg)	h'' (kJ/kg)	$\Delta h_v = h'' - h'$ (kJ/kg)	s' (kJ/(kgK))	s'' (kJ/(kgK))
−110	0.00730	0.62591	21 441.0	79.474	354.05	274.57	0.43930	2.1222
−100	0.01991	0.63636	8338.8	90.056	358.80	268.75	0.50224	2.0544
−90	0.04778	0.64725	3667.5	100.65	363.64	262.98	0.56174	1.9976
−80	0.10319	0.65866	1785.5	111.29	368.53	257.24	0.61824	1.9501
−70	0.20398	0.67064	945.76	121.97	373.44	251.47	0.67241	1.9100
−60	0.37425	0.68329	537.47	132.73	378.34	245.61	0.72377	1.8761
−50	0.64457	0.69669	323.97	143.58	383.18	239.60	0.77342	1.8472
−40	1.0519	0.71096	205.18	154.54	387.92	233.38	0.82134	1.8223
−30	1.6389	0.72626	135.46	165.63	392.52	226.88	0.86776	1.8009
−20	2.4538	0.74275	92.621	176.89	396.92	220.03	0.91288	1.7821
−10	3.5492	0.76065	65.224	188.33	401.09	212.76	0.95690	1.7654
0	4.9817	0.78027	47.078	200.00	404.98	204.98	1.0000	1.7505
10	6.8115	0.80196	34.684	211.93	408.52	196.60	1.0424	1.7367
20	9.1018	0.82623	25.983	224.16	411.65	187.50	1.0842	1.7238
30	11.919	0.85380	19.721	236.76	414.29	177.53	1.1256	1.7112
40	15.334	0.88571	15.109	249.80	416.30	166.50	1.1670	1.6987
50	19.421	0.92360	11.638	263.41	417.51	154.10	1.2086	1.6855
60	24.265	0.97028	8.9656	277.78	417.65	139.87	1.2510	1.6708
70	29.957	1.0312	6.8541	293.24	416.20	122.96	1.2950	1.6534
80	36.616	1.1195	5.1213	310.52	412.11	101.60	1.3426	1.6303
90	44.404	1.2827	3.5651	331.97	401.92	69.945	1.3999	1.5925

Reference points: see footnote of Table 4.10

each other. The triple point of water is 273.16 K by definition, which corresponds to a pressure at the triple point of 611.657 Pa.

Caloric Properties

During the freezing of a liquid the latent heat of fusion Δh_f is released (Table 4.15). At the same time the liquid entropy is reduced by $\Delta s_f = \Delta h_f / T_f$ with T_f being the melting or freezing temperature. According to the Dulong–Petit law the molar specific heat divided by the number of atoms in the molecule is, above ambient temperature, about 25.9 kJ/kmol K. If absolute zero is approached, this approximation rule is no longer valid. Therefore, the molar specific heat at constant volume is for all solids

$$\overline{C} = a(T/\Theta)^3, \quad \text{for} \quad T/\Theta < 0.1$$

with $a = 4782.5$ J/mol K and where Θ is the Debye temperature (Table 4.16).

4.6.5 Mixing Temperature. Measurement of Specific Heats

If several substances with different masses m_i, temperatures t_i, and specific heats c_{pi} ($i = 1, 2, \ldots$) are mixed at constant pressure without external heat supply, a mixing temperature t_m arises after a sufficient period of time. It is

$$t_m = \left(\sum m_i c_{pi} t_i\right) / \left(\sum m_i c_{pi}\right)$$

with c_{pi} being the mean specific heats between 0 °C and t °C. It is possible to calculate an unknown specific heat from the measured temperature t_m, if all other specific heats are known.

Table 4.14 Thermal extension $(l - l_0)/l_0$ of some solids in mm/m in the temperature interval between 0 °C and t °C; l_0 is the length at 0 °C

Substance	0 to −190 °C	0 to 100 °C	0 to 200 °C	0 to 300 °C	0 to 400 °C	0 to 500 °C	0 to 600 °C	0 to 700 °C	0 to 800 °C	0 to 900 °C	0 to 1000 °C
Aluminium	−3.43	2.38	4.90	7.65	10.60	13.70	17.00	–	–	–	–
Lead	−5.08	2.90	5.93	9.33	–	–	–	–	–	–	–
Al-Cu-Mg [0.95 Al; 0.04 Cu + Mg, Mn, St, Fe]	–	2.35	4.90	7.80	10.70	13.65	–	–	–	–	–
Iron–nickel alloy [0.64 Fe; 0.36 Ni]	–	0.15	0.75	1.60	3.10	4.70	6.50	8.5	10.5	12.55	–
Iron–nickel alloy [0.77 Fe; 0.23 Ni]	–	–	–	2.80	4.00	5.25	6.50	7.80	9.25	10.50	11.85
Glass: Jena, 16 III	−1.13	0.81	1.67	2.60	3.59	4.63	–	–	–	–	–
Glass: Jena, 1565 III	–	0.345	0.72	1.12	1.56	2.02	–	–	–	–	–
Gold	−2.48	1.42	2.92	4.44	6.01	7.62	9.35	11.15	13.00	14.90	–
Gray cast iron	−1.59	1.04	2.21	3.49	4.90	6.44	8.09	9.87	11.76	–	–
Constantane [0.60 Cu; 0.40 Ni]	−2.26	1.52	3.12	4.81	6.57	8.41	–	–	–	–	–
Copper	−2.65	1.65	3.38	5.15	7.07	9.04	11.09	–	–	–	–
Sintered magnesia	–	–	2.45	3.60	4.90	6.30	7.75	9.30	10.80	12.35	13.90
Magnesium	−4.01	2.60	5.41	8.36	11.53	14.88	–	–	–	–	–
Manganese bronze [0.85 Cu; 0.09 Mn; 0.06 Sn]	−2.84	1.75	3.58	5.50	7.51	9.61	–	–	–	–	–
Manganin [0.84 Cu; 0.12 Mn; 0.04 Ni]	–	1.75	3.65	5.60	7.55	9.70	11.90	14.3	16.80	–	–
Brass [0.62 Cu; 0.38 Zn]	−3.11	1.84	3.85	6.03	8.39	–	–	–	–	–	–
Molybdenum	−0.79	0.52	1.07	1.64	2.24	–	–	–	–	–	–
Nickel	−1.89	1.30	2.75	4.30	5.95	7.60	9.27	11.05	12.89	14.80	16.80
Palladium	−1.93	1.19	2.42	3.70	5.02	6.38	7.79	9.24	10.74	12.27	13.86
Platinum	−1.51	0.90	1.83	2.78	3.76	4.77	5.80	6.86	7.94	9.05	10.19
Platinum-iridium-alloy [0.80 Pt; 0.20 Ir]	−1.43	0.83	1.70	2.59	3.51	4.45	5.43	6.43	7.47	8.53	9.62
Quartz glass	+0.03	0.05	0.12	0.19	0.25	0.31	0.36	0.40	0.45	0.50	0.54
Silver	−3.22	1.95	4.00	6.08	8.23	10.43	12.70	15.15	17.65	–	–
Sintered corundum	–	–	1.30	2.00	2.75	3.60	4.45	5.30	6.25	7.15	8.15
Steel, soft	−1.67	1.20	2.51	3.92	5.44	7.06	8.79	10.63	–	–	–
Steel, hard	−1.64	1.17	2.45	3.83	5.31	6.91	8.60	10.40	–	–	–
Zinc	−1.85	1.65	–	–	–	–	–	–	–	–	–
Tin	−4.24	2.67	–	–	–	–	–	–	–	–	–
Tungsten	−0.73	0.45	0.90	1.40	1.90	2.25	2.70	3.15	3.60	4.05	4.60

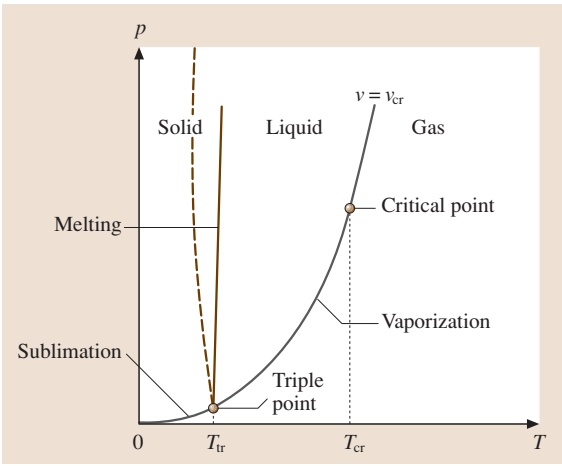


Fig. 4.11 *p–T* diagram with the three boundary lines of the phases. (The gradient of the melting line of water is negative, *dashed line*)

Example 4.8: A mass of $m_a = 0.2\,\mathrm{kg}$ aluminium at $t_a = 100\,^\circ\mathrm{C}$ is inserted into a thermally perfectly isolated calorimeter, which is filled with 0.8 kg water ($c_p = 4.186\,\mathrm{kJ/kg\,K}$) at 15 °C and which consists of 0.25 kg silver ($c_{ps} = 0.234\,\mathrm{kJ/kg\,K}$). After reaching equilibrium a mixing temperature of 19.24°*C* is measured. What is the specific heat of aluminium? Therefore,

$$t_m = \frac{(mc_p + m_s c_{ps})t + m_a c_{pa} t_a}{mc_p + m_s c_{ps} + m_a c_{pa}} ,$$

which resolves to

$$c_{pa} = \frac{(mc_p + m_s c_{ps})(t - t_m)}{m_a (t_m - t_a)} ,$$

$$c_{pa} = \left(0.8\,\mathrm{kg} \times 4.186 \frac{\mathrm{kJ}}{\mathrm{kgK}} + 0.25\,\mathrm{kg} \times 0.234 \frac{\mathrm{kJ}}{\mathrm{kgK}}\right) \times \frac{15\,^\circ\mathrm{C} - 19.24\,^\circ\mathrm{C}}{0.2\,\mathrm{kg}(19.24\,^\circ\mathrm{C} - 100\,^\circ\mathrm{C})} .$$

$$c_{pa} = 0.894\,\mathrm{kJ/kg\,K} .$$

4.7 Changes of State of Gases and Vapors

4.7.1 Change of State of Gases and Vapors in Closed Systems

A closed thermodynamic system has a fixed mass Δm. The following changes of state (constant volume, constant pressure, and constant temperature) are idealized limiting cases of real changes of state. The gas volume remains unchanged during changes of state at *constant volume* or *isochoric* changes of state, e.g., when a gas volume is in a container with solid, rigid walls. No work is done, and the supplied heat transfer causes only a change of internal energy. During a change of state at *constant pressure* or an *isobaric* change of state the gas volume has to increase if heat transfer is supplied. The supplied heat transfer increases the enthalpy during reversible changes of state. Changes of state at constant temperature are also called isothermal changes of state. Apart from a very few exceptions, heat has to be transferred to the gas during expansion and transferred from the gas during compression in order for the temperature to remain constant. For an ideal gas at constant temperature, the internal energy does not change, as $U(T) = \text{const.}$, thus the supplied heat transfer is equal to the work done by the system. The isotherm of an ideal gas ($pV = mRT = \text{const.}$) is a hyperbola in the *p–V* diagram.

During adiabatic changes of state there is no heat exchange between the system and its environment. The cases are approximately realized in compressors and expansion machines, because there the compression or expansion takes place in a very short period of time such that little heat transfer is exchanged with the environment during the change of state. According to the second law (see Sect. 4.4.3) the entire entropy change is caused by internal irreversibilities of the system, $\dot{S} = \dot{S}_{gen}$. A reversible adiabatic process proceeds at constant entropy $\dot{S} = 0$, where such a change of state is called isentropic. Thus, a reversible adiabatic process is at the same time also an isentropic process. However, an isentropic process is not necessarily also an adiabatic process, because from $\dot{S} = \dot{S}_Q + \dot{S}_{gen} = 0$ it does not always follow that $\dot{S}_Q = 0$.

Figure 4.13 shows the different changes of state in *p–V* and *T–S* diagrams. Additionally, the most important relationships for the properties of ideal gases are given.

An isothermal change of state requires perfect heat exchange, whereas no heat exchange at all with the environment must take place during an adiabatic change of state. Both cannot be achieved completely in reality. Therefore, a *polytropic* change of state is introduced via

$$pV^n = \text{const.} , \qquad (4.89)$$

Table 4.15 Thermal engineering properties: density ρ, specific heat c_p for 0–100 °C, melting temperature t_f, latent heat of fusion Δh_f, boiling temperature t_s and enthalpy of vaporization Δh_v

	ρ (kg/dm^3)	c_p (kJ/(kgK))	t_f (°C)	Δh_f (kJ/kg)	t_s (°C)	Δh_v (kJ/kg)
Solids (metals and sulfur) at 1.0132 bar						
Aluminium	2.70	0.921	660	355.9	2270	11 723
Antimony	6.69	0.209	630.5	167.5	1635	1256
Lead	11.34	0.130	327.3	23.9	1730	921
Chrome	7.19	0.506	1890	293.1	2642	6155
Iron (pure)	7.87	0.465	1530	272.1	2500	6364
Gold	19.32	0.130	1063	67.0	2700	1758
Iridium	22.42	0.134	2454	117.2	2454	3894
Copper	8.96	0.385	1083	209.3	2330	4647
Magnesium	1.74	1.034	650	209.3	1100	5652
Manganese	7.3	0.507	1250	251.2	2100	4187
Molybdenum	10.2	0.271	2625	–	3560	7118
Nickel	8.90	0.444	1455	293.1	3000	6197
Platinum	21.45	0.134	1773	113.0	3804	2512
Mercury	13.55	0.138	−38.9	11.7	357	301
Silver	10.45	0.234	960.8	104.7	1950	2177
Titanium	4.54	0.471	1800	–	3000	–
Bismuth	9.80	0.126	271	54.4	1560	837
Tungsten	19.3	0.134	3380	251.2	6000	4815
Zinc	7.14	0.385	419.4	112.2	907	1800
Tin	7.28	0.226	231.9	58.6	2300	2596
Sulfur (rhombic)	2.07	0.720	112.8	39.4	444.6	293
Liquids at 1.0132 bar						
Ethyl alcohol	0.79	2.470	−114.5	104.7	78.3	841.6
Ethyl ether	0.71	2.328	−116.3	100.5	34.5	360.1
Acetone	0.79	2.160	−94.3	96.3	56.1	523.4
Benzene	0.88	1.738	5.5	127.3	80.1	395.7
Glycerin [a]	1.26	2.428	18.0	200.5	290.0	854.1
Saline solution (saturated)	1.19	3.266	−18.0	–	108.0	–
Sea water (3.5% salt content)	1.03	–	−2.0	–	100.5	–
Methyl alcohol	0.79	2.470	−98.0	100.5	64.5	1101.1
n-Heptane	0.68	2.219	−90.6	141.5	98.4	318.2
n-Hexane	0.66	1.884	−95.3	146.5	68.7	330.8
Spirits of turpentine	0.87	1.800	−10.0	116.0	160.0	293.1
Water	1.00	4.183	0.0	333.5	100.0	2257.1

[a] Solidification point at 0 °C. Melting and solidification point do not always coincide

whereas in practice n is usually between 1 and $\varkappa$. Isochore, isobar, isotherm, and reversible adiabate are special cases of a polytrope with the following exponents (Fig. 4.12): isochore ($n = \infty$), isotherm ($n = 1$),

Table 4.15 (cont.)

	ρ (kg/dm^3)	c_p (kJ/(kgK))	t_f (°C)	Δh_f (kJ/kg)	t_s (°C)	Δh_v (kJ/kg)
Gases at 1.0132 bar and 0 °C						
Ammonia	0.771	2.060	−77.7	332.0	−33.4	1371
Argon	1.784	0.523	−189.4	29.3	−185.9	163
Ethylene	1.261	1.465	−169.5	104.3	−103.9	523
Helium	0.178	5.234	–	37.7	−268.9	21
Carbon dioxide	1.977	0.825	−56.6	180.9	−78.5 [b]	574
Carbon oxide	1250	1.051	−205.1	30.1	−191.5	216
Air	1.293	1.001	–	–	−194.0	197
Methane	0.717	2.177	−182.5	58.6	−161.5	548
Oxygen	1.429	0.913	−218.8	13.8	−183.0	214
Sulfur dioxide	2.926	0.632	−75.5	115.6	−10.2	390
Nitrogen	1.250	1.043	−210.0	25.5	−195.8	198
Hydrogen	0.09	14.235	−259.2	58.2	−252.8	454

[b] CO_2 does not boil, but sublimates at 1.0132 bar

Table 4.16 Debye temperatures of some substances

Metal	Θ (K)	Other substances	Θ (K)
Pb	88	KBr	177
Hg	97	KCl	230
Cd	168	NaCl	281
Na	172	C	1860
Ag	215		
Ca	226		
Zn	235		
Cu	315		
Al	398		
Fe	453		

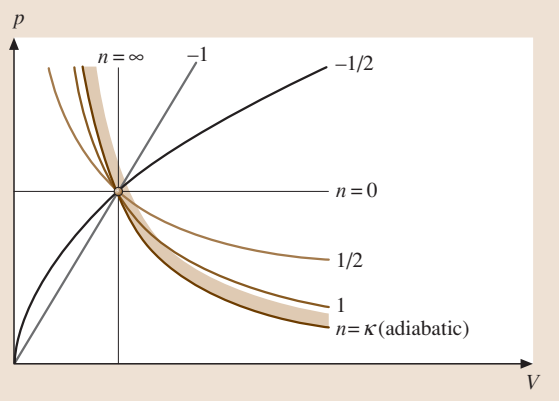


Fig. 4.12 Polytropic processes with different exponents

and reversible adiabate ($n = \varkappa$). It holds further that

$$
\begin{aligned}
v_2/v_1 &= (p_1/p_2)^{1/n} = (T_1/T_2)^{1/(n-1)} \,, \\
W_{12} &= mR(T_2 - T_1)/(n-1) \\
&= (p_2V_2 - p_1V_1)/(n-1) \\
&= p_1V_1\left[(p_2/p_1)^{(n-1)/n} - 1\right]/(n-1) \qquad (4.90)
\end{aligned}
$$

and

$$
W_{t12} = nW_{12} \,. \qquad (4.91)
$$

The heat exchanged is

$$
Q_{12} = mc_v(n - \varkappa)(T_2 - T_1)/(n-1) \,. \qquad (4.92)
$$

Example 4.9: A compressed air system should deliver 1000 m^3_n compressed air of 15 bar per hour (note: 1 m^3_n = 1 standard cubic meter for a gas volume at 0 °C and 1.01325 bar). The air inlet is at a pressure of $p_1 = 1$ bar and a temperature $t_1 = 20$ °C. The adiabatic exponent of air is $\varkappa = 1.4$. How much power is required, if the compression is polytropic with $n = 1.3$? What heat transfer rate has to be exchanged for this process?

The inlet air volume flow is, as given in the requirements, 1000 m^3 at 0 °C and 1.01325 bar,

$$
\begin{aligned}
\dot{V}_1 &= \frac{p_0T_1}{p_1T_0}\dot{V}_0 \\
&= \frac{1.01325 \times 293.15}{1 \times 273.15} 1000 \, \frac{\mathrm{m}^3}{\mathrm{h}} \\
&= 1087.44 \, \frac{\mathrm{m}^3}{\mathrm{h}} \,.
\end{aligned}
$$

For polytropic changes of state, Eqs. (4.91) and (4.90) yield

$$P = \dot{W}_t = \frac{n p_1 \dot{V}_1}{n-1}\left[\left(\frac{p_2}{p_1}\right)^{\frac{n-1}{n}} - 1\right]$$
$$= \frac{1.3 \times 10^5 \frac{\mathrm{N}}{\mathrm{m}^2} 1087.44 \frac{\mathrm{m}^3}{\mathrm{h}}}{1.3-1}\left(15^{\frac{1.3-1}{1.3}} - 1\right)$$
$$= 113.6\,\mathrm{kW}\,.$$

According to (4.91) and (4.92),

$$\frac{Q_{12}}{W_{t12}} = \frac{\dot{Q}}{P} = c_v \frac{n-\varkappa}{nR}$$

or, since $R = c_p - c_v$ and $\varkappa = c_p/c_v$

$$\frac{\dot{Q}}{P} = \frac{1}{n}\frac{n-\varkappa}{\varkappa-1}\,.$$

Thus, $\dot{Q} = \frac{1}{1.3}\frac{1.3-1.4}{1.4-1} 113.6\,\mathrm{kW} = -21.85\,\mathrm{kW}$.

4.7.2 Changes of State of Flowing Gases and Vapors

In order to describe the flow of a fluid mass Δm, in addition to the thermodynamic properties, the size and direction of the velocity everywhere in the field are also required. The following discussion is limited to steady flows in channels with constant, diverging, or converging cross sections.

In addition to the first and the second law the conservation of mass law holds

$$\dot{m} = A w \varrho = \text{const.} \qquad (4.93)$$

For a flow that does no work on the environment the first law, (4.21), is reduced to

$$\Delta m(h_2 - h_1) + \Delta m\left(\frac{w_2^2}{2} - \frac{w_1^2}{2}\right) + \Delta m g(z_2 - z_1)$$
$$= Q_{12}\,, \qquad (4.94)$$

regardless of whether the flow is reversible or irreversible. Neglecting changes in potential energy, it holds that for an adiabatic flow

$$h_2 - h_1 + \frac{w_2^2}{2} - \frac{w_1^2}{2} = 0\,. \qquad (4.95)$$

Thus, the increase in kinetic energy is equal to the decrease in enthalpy of the fluid. For an adiabatic throttle process, it follows from (4.93), provided A, $\varrho = \text{const.}$, that $w = \text{const.}$ and thus from (4.95) that $h_1 = h_2 = \text{const}$. The pressure reduction in an adiabatic throttle process is accompanied by an entropy increase, since the process is irreversible. According to (4.32), the enthalpy change in a reversible adiabatic flow is caused by a change in pressure, $\mathrm{d}h = v\,\mathrm{d}p$.

Flow of Ideal Gases

Applying (4.95) to an ideal gas exiting a vessel (Fig. 4.14), in which the gas in the vessel possesses

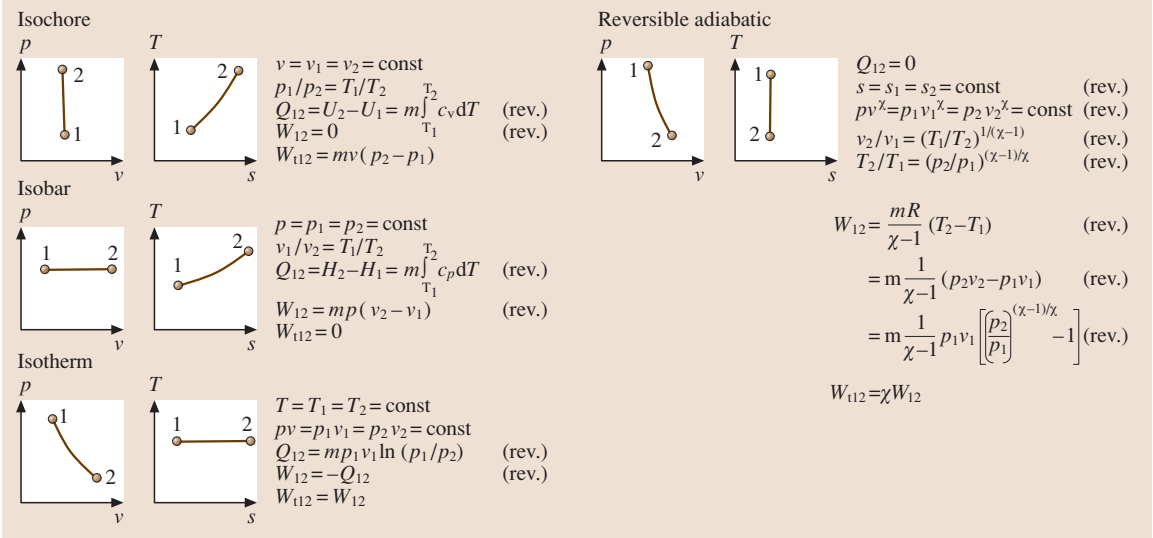


Fig. 4.13 Changes of state of ideal gases. The (rev.) denotes that the change of state is reversible.

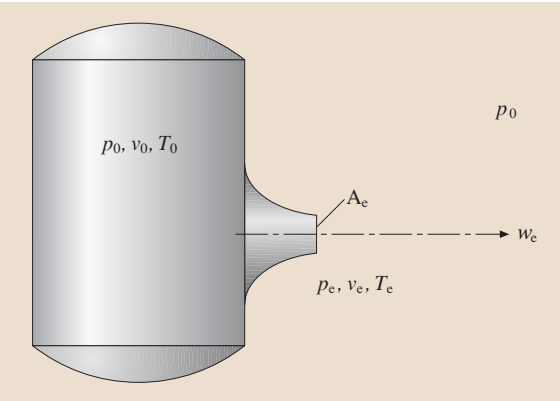


Fig. 4.14 Flow out of a pressure vessel

the constant state p_0, v_0, T_0 with $w_0 = 0$, and where $h_e - h_0 = c_p(T_e - T_0)$, $w_0 = 0$, leads to

$$\frac{w_e^2}{2} = c_p(T_0 - T_e) = c_p T_0\left(1 - \frac{T_e}{T_0}\right).$$

For a reversible adiabatic change of state, according to (4.90), $T_e/T_0 = (p_e/p_0)^{(\varkappa-1)/\varkappa}$. Additionally, it holds that $T_0 = p_0 v_0/R$ according to (4.60) and $c_p/R = \varkappa/(\varkappa - 1)$ according to (4.73). Thus, the exit velocity is given by

$$w_e = \sqrt{2\frac{\varkappa}{\varkappa - 1} p_0 v_0 \left[1 - \left(\frac{p_e}{p_0}\right)^{\frac{\varkappa-1}{\varkappa}}\right]}. \qquad (4.96)$$

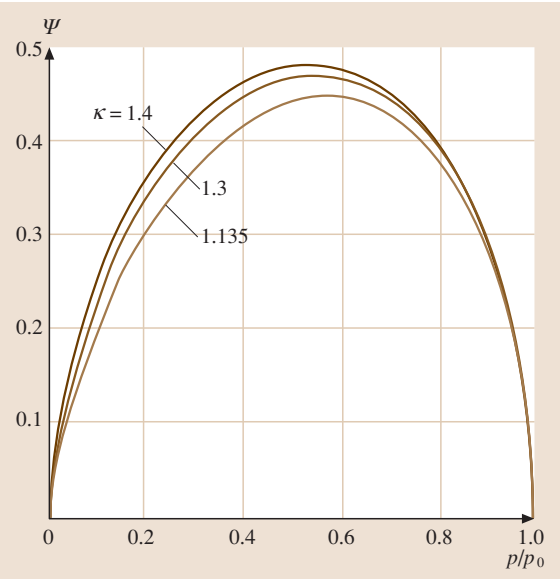


Fig. 4.15 Outlet function Ψ

Taking into account $p_0 v_0^* = p_e v_e^*$, the out-flowing mass $\dot{m} = A_e w_e / v_e$ is

$$\dot{m} = A\Psi\sqrt{2p_0/v_0} \qquad (4.97)$$

with the outlet function

$$\Psi = \sqrt{\frac{\varkappa}{\varkappa - 1}}\sqrt{\left(\frac{p}{p_0}\right)^{\frac{2}{\varkappa}} - \left(\frac{p}{p_0}\right)^{\frac{\varkappa+1}{\varkappa}}}. \qquad (4.98)$$

The result is a function of the adiabatic exponent $\varkappa$ and of the pressure ratio p/p_0 (Fig. 4.15) and has a maximum Ψ_{max}, which can be determined from evaluating $d\Psi/d(p/p_0) = 0$. This maximum corresponds to a specific pressure ratio that is called the *Laval pressure ratio*

$$\frac{p_s}{p_0} = \left(\frac{2}{\varkappa + 1}\right)^{\frac{\varkappa}{\varkappa-1}}. \qquad (4.99)$$

At this pressure ratio

$$\Psi_{max} = \left(\frac{2}{\varkappa + 1}\right)^{\frac{1}{\varkappa-1}}\sqrt{\frac{\varkappa}{\varkappa + 1}}. \qquad (4.100)$$

Corresponding to this pressure ratio, according to (4.96) with $p_e/p_0 = p_s/p_0$ and a velocity $w_e = w_s$, is the relation

$$w_s = \sqrt{2\frac{\varkappa}{\varkappa + 1} p_0 v_0} = \sqrt{\varkappa p_s v_s} = \sqrt{\varkappa R T_s}. \qquad (4.101)$$

This is equal to the sonic velocity in state p_s, v_s. Generally, the sonic velocity is the velocity at which pressure and density fluctuations are transmitted. For reversible adiabatic changes of state it is given by

$$w_s = \sqrt{(\partial p/\partial \varrho)_s}.$$

Thus, for ideal gases it takes on the value $w_s = \sqrt{\varkappa RT}$, where the sonic velocity is a property.

Example 4.10: A steam boiler produces 10 t of saturated vapor at $p_0 = 15$ bar. The vapor may be treated as an ideal gas ($\varkappa = 1.3$). How large must the cross section of the safety relief valve be in order to be able to discharge the entire vapor mass flow?

Since the out-flowing mass $\dot{m}$ is constant in every cross section, it follows from (4.97) that $A\Psi = \text{const.}$ as well. As the discharge flow area is decreased, A decreases, and Ψ increases, reaching at most the value Ψ_{max}. Then the back pressure is less than or equal to the Laval pressure. In the present case the back pressure of the atmosphere, $p = 1$ bar, is less than the Laval pressure, which is

Table 4.17 Composition and calorific values of solid fuels

Fuel	Ash (mass%)	Water (mass%)	Composition of ash-free dry substance (mass%)					Gross calorific value (MJ/kg)	Net calorific value (MJ/kg)
			C	H	S	O	N		
Wood, air-dried	< 0.5	10–20	50	6	0.0	43.9	0.1	15.91–18.0	14.65–16.75
Peat, air-dried	< 15	15–35	50–60	4.5–6	0.3–2.5	30–40	1–4	13.82–16.33	11.72–15.07
Raw soft coal	2–8	50–60	65–75	5–8	0.5–4	15–26	0.5–2	10.47–12.98	8.37–11.30
Soft coal briquette	3–10	12–18						20.93–21.35	19.68–20.10
Hard coal	3–12	0–10	80–90	4–9	0.7–1.4	4–12	0.6–2	29.31–35.17	27.31–34.12
Anthracite	2–6	0–5	90–94	3–4	0.7–1	0.5–4	1–1.5	33.49–34.75	32.66–33.91

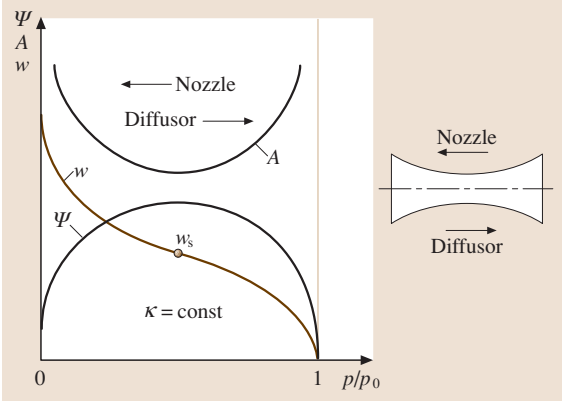


Fig. 4.16 Nozzle and diffusor flow

calculated with (4.99) to be 8.186 bar. With this result the required cross section follows from (4.97), if $\Psi = \Psi_{max} = 0.472$ according to (4.100) is inserted. With $\dot{m} = 10 \times 10^3 \times (1/3600)\,kg/s = 2.7778\,kg/s$ and $v_0 = v'' = 0.1317\,m^3/kg$ (according to Table 4.9 at $p_0 = 15$ bar) it follows from (4.97) that $A = 12.33\,cm^2$. Because of the jet's contraction, where the size depends on the design of the valve, an increase should be added.

Jet and Diffusion Flow

As shown in Fig. 4.16, for a given adiabatic exponent $\varkappa$ a certain pressure ratio p/p_0 corresponds to a specific value of the outlet function Ψ. Since the mass flow $\dot{m}$ is constant in each cross section, it follows from (4.97) that also $A\Psi = \text{const}$. Thus, it is possible to assign to each pressure ratio a certain cross section A; see Fig. 4.16. Two cases have to be distinguished:

a) The pressure decreases in the flow direction. The curves Ψ, A, and w in Fig. 4.16 are passed through from right to left. At first the cross section A decreases, then it increases again. The velocity increases and goes from subsonic to supersonic. The kinetic energy of the flow increases. Such an apparatus is called a nozzle. In a nozzle that operates only in the subsonic regime the cross section decreases continuously, whereas it increases in the supersonic regime. In a nozzle narrowing in the flow direction the pressure at the outlet cross section cannot decrease below the Laval pressure, even if the outside pressure is arbitrarily small. This follows from $A\Psi = \text{const}$. Since A decreases in the flow direction, Ψ can only increase, reaching at most the value Ψ_{max} to which the Laval pressure ratio corresponds.
If the pressure at the outlet cross section of a nozzle is reduced below the pressure value corresponding to the outlet cross section, the jet expands after leaving the nozzle. If the back pressure is increased above the correct value, the pressure increase moves upstream where in this case the gas exits with subsonic velocity. If the gas exits with sonic, or in a diverging nozzle with supersonic velocity, a shock occurs at the nozzle outlet and the pressure increases to the pressure of the environment.

b) The pressure increases in the flow direction. The curves Ψ, A, and w in Fig. 4.16 are passed through from left to right. At first the cross section decreases, then increases again. The velocity decreases from supersonic to subsonic, and the kinetic energy decreases while the pressure increases. Such an apparatus is called a diffuser. In a diffuser that works only in the subsonic regime the cross section increases continuously, whereas it decreases in the supersonic regime.

4.8 Thermodynamic Processes

4.8.1 Combustion Processes

Heat transfer for technical processes is still mostly obtained through combustion. Combustion is a chemical reaction during which a substance, e.g., carbon, hydrogen, or hydrocarbons, is oxidized and which is strongly exothermic, i. e., a large quantity of heat is released. Fuels can be solid, liquid, or gaseous. The required oxygen is mostly provided by atmospheric air. To start a combustion process the fuel must be brought above its ignition temperature, which, in turn, varies according to the type of fuel being used. The main components of all important technical fuels are carbon C, and hydrogen H. In addition, oxygen O, and, with the exception of natural gas, a certain amount of sulfur are also present. Sulfur reacts during a combustion process to produce the unwanted compound sulfur dioxide SO_2.

Equations of Reactions

The elements H, C, and S, which are contained in fuels as mentioned above, are burned to CO_2, H_2O, and SO_2, if complete combustion takes place. The equation of reaction leads to the required amount of oxygen and to the resulting amount of each product in the exhaust gas. For the combustion of carbon C it holds that

$$\begin{aligned} \mathrm{C} + \mathrm{O_2} &= \mathrm{CO_2}\,, \\ 1\,\mathrm{kmol\ C} + 1\,\mathrm{kmol\ O_2} &= 1\,\mathrm{kmol\ CO_2}\,, \\ 12\,\mathrm{kg\ C} + 32\,\mathrm{kg\ O_2} &= 44\,\mathrm{kg\ CO_2}\,. \end{aligned}$$

From this it follows that the minimum oxygen demand for complete combustion is

$$o_{\min} = 1/12\,\mathrm{kmol/kg\ C}$$

or

$$O_{\min} = 1\,\mathrm{kmol/kmol\ C}.$$

The minimum air demand for complete combustion is called the *theoretical air* and results from the oxygen fraction of 21 mol% in air

$$l_{\min} = (o_{\min}/0.21)\ \mathrm{kmol\ air\,/\,kg\ C}$$

or

$$L_{\min} = (O_{\min}/0.21)\ \mathrm{kmol\ air\,/\,kmol\ C}\,.$$

The amount of CO_2 in the exhaust gas is $(1/12)$ kmol/kg C. Similarly, the equations of reaction for the combustion of hydrogen H_2 and sulfur S are

$$\begin{aligned} \mathrm{H_2} + 1/2\,\mathrm{O_2} &= \mathrm{H_2O}\,, \\ 1\,\mathrm{kmol\ H_2} + 1/2\,\mathrm{kmol\ O_2} &= 1\,\mathrm{kmol\ H_2O}\,, \\ 2\,\mathrm{kg\ H_2} + 16\,\mathrm{kg\ O_2} &= 18\,\mathrm{kg\ H_2O}\,, \end{aligned}$$

$$\begin{aligned} \mathrm{S} + \mathrm{O_2} &= \mathrm{SO_2}\,, \\ 1\,\mathrm{kmol\ S} + 1\,\mathrm{kmol\ O_2} &= 1\,\mathrm{kmol\ SO_2}\,, \\ 32\,\mathrm{kg\ S} + 32\,\mathrm{kg\ O_2} &= 64\,\mathrm{kg\ SO_2}\,. \end{aligned}$$

Denoting the carbon, hydrogen, sulfur, and oxygen fractions by c, h, s, and o in kg per kg fuel, according to the above calculations, the minimum oxygen demand becomes

$$o_{\min} = \left(\frac{c}{12} + \frac{h}{4} + \frac{s}{32} - \frac{o}{32}\right)\mathrm{kmol/kg}\,, \qquad (4.102)$$

or for short

$$o_{\min}\frac{1}{12}c\sigma\ \mathrm{kmol/kg}\,, \qquad (4.103)$$

where σ is a characteristic of the fuel (O_2 demand in kmol related to the kmol C in the fuel). The actual air demand (related to 1 kg fuel) is

$$l = \lambda l_{\min} = (\lambda o_{\min}/0.21)\ \mathrm{kmol\ air/kg}\,, \qquad (4.104)$$

where λ is the excess air number. In addition to the combustion products CO_2, H_2O, and SO_2, exhaust gases also ordinarily contain water with a content of $w/18$ (SI units of kmol per kg fuel), and the supplied combustion air l less the spent oxygen $o_{\min}$. The supplied combustion air is therefore assumed to be dry or it is assumed that the water vapor content is negligibly small. The following exhaust amounts, related to 1 kg of fuel, are given by

$$\begin{aligned} n_{\mathrm{CO_2}} &= c/12\,, \\ n_{\mathrm{H_2O}} &= h/2 + w/18\,, \\ n_{\mathrm{SO_2}} &= s/32\,, \\ n_{\mathrm{O_2}} &= (\lambda - 1)o_{\min}\,, \\ n_{\mathrm{N_2}} &= 0.791\,. \end{aligned}$$

The sum is the total exhaust amount

$$n_{\mathrm{exh}} = \left[c/12 + h/2 + w/18 + s/32 + (\lambda - 1)o_{\min} + 0.791\right]\mathrm{kmol/kg}.$$

Table 4.18 Net calorific values of the simplest fuels at 25 °C and 1.01325 bar

	C	CO	H_2 (gross calorific value)	H_2 (net calorific value)	S
kJ/kmol	393 510	282 989	285 840	241 840	296 900
kJ/kg	32 762	10 103	141 800	119 972	9260

This can be simplified by using (4.102) and (4.104) to yield

$$n_{exh} = \left[l + \frac{1}{12}\left(3h + \frac{3}{8}o + \frac{2}{3}w\right)\right] \text{ kmol/kg} . \tag{4.105}$$

Example 4.11: 500 kg coal with the composition $c = 0.78$, $h = 0.05$, $o = 0.08$, $s = 0.01$, and $w = 0.02$ and an ash content $a = 0.06$ are completely burned per hour in a furnace with excess air number $\lambda = 1.4$. How much air is necessary, how much exhaust arises, and what is its composition?

The minimum oxygen demand is determined according to (4.102)

$$o_{min} = \left(\frac{0.78}{12} + \frac{0.05}{4} + \frac{0.01}{32} - \frac{0.08}{32}\right) \text{kmol/kg}$$
$$= 0.0753 \text{ kmol/kg} .$$

The minimum air demand is

$$l_{min} = o_{min}/0.21 = 0.3586 \text{ kmol/kg} .$$

The amount of air that has to be supplied is

$$l = \lambda l_{min} = 1.4 \times 0.3586 = 0.502 \text{ kmol/kg} .$$

Thus $0.502 \text{ kmol/kg} \times 500 \text{ kg/h} = 251 \text{ kmol/h}$. With the molar mass of air $M = 28.953$ kg/kmol, the air demand becomes $0.502 \times 28.953 \text{ kg/kg} = 14.54 \text{ kg/kg}$. Thus, $14.54 \text{ kg/kg} \times 500 \text{ kg/h} = 7270 \text{ kg/h}$. The exhaust amount is determined according to (4.105)

$$n_{exh} = (0.502 + 1/12(3 \times 0.05 + 3/8 \times 0.08$$
$$+ 2/3 \times 0.02)) \text{ kmol/kg}$$
$$= 0.518 \text{ kmol/kg} .$$

Thus $0.581 \text{ kmol/kg} \times 500 \text{ kg/h} = 259 \text{ kmol/h}$ with 0.065 kmol CO_2/kg, 0.0261 kmol H_2O/kg, 0.0003 kmol SO_2/kg, 0.3966 kmol N_2/kg and 0.0301 kmol O_2/kg.

Net Calorific Value and Gross Calorific Value

The net calorific value is the energy released during combustion, if the exhaust gases are cooled down to the temperature at which the fuel and air are supplied. Water is included in the exhaust gases as vapor. If the water vapor is condensed, the released heat is called the gross calorific value. Net and gross calorific values are specified, according to DIN 51900, for combustion at atmospheric pressure, if all involved substances possess a temperature of 25 °C before and after combustion. Net and gross calorific values (Tables 4.18–4.20) are independent of the amount of excess air and are only a characteristic of the fuel. The gross calorific value Δh_{gcv} exceeds the net calorific value Δh_{ncv} by the enthalpy of vaporization Δh_v of the water included in the exhaust gas

$$\Delta h_{gcv} = \Delta h_{ncv} + (8.937h + w)\, \Delta h_v .$$

Because the water leaves technical furnaces mostly as vapor, often only the net calorific value can be utilized. The net calorific value of heating oil can be expressed quite well, as experience shows [4.19], by the equation

$$\Delta h_{ncv} = (54.04 - 13.29\varrho - 29.31s) \text{ MJ/kg} , \tag{4.106}$$

where the density ϱ of the heating oil in kg/dm^3 is at 15 °C and the sulfur content s is in kg/kg.

Example 4.12: What is the net calorific value of a light heating oil with a density of $\varrho = 0.86$ kg/dm^3 and a sulfur content of $s = 0.8$ mass%? According to (4.106)

$$\Delta h_{ncv} = (54.04 - 13.29 \times 0.86$$
$$- 29.31 \times 0.8 \times 10^{-2}) \text{ MJ/kg}$$
$$= 42.38 \text{ MJ/kg} .$$

Combustion Temperature

The theoretical combustion temperature is the temperature of the exhaust gas at complete isobar-adiabatic combustion if no dissociation takes place. The heat released during combustion increases the internal energy and thus the temperature of the gas, which provides the basis for doing flow work. The theoretical combustion temperature is calculated under the condition that the enthalpy of all substances transferred to the combustion

Table 4.19 Combustion of liquid fuels

Fuel	Molar weight	Content (mass%)		Characteristic	Calorific value (kJ/kg)	
	(kg/kmol)	C	H	σ	Gross	Net
Ethyl alcohol C_2H_5OH	46.069	52	13	1.50	29 730	26 960
Spirit 95%	–	–	–	1.50	28 220	25 290
90%	–	–	–	1.50	26 750	23 860
85%	–	–	–	1.50	25 250	22 360
Benzene (pure) C_6H_6	78.113	92.2	7.8	1.25	41 870	40 150
Toluene (pure) C_7H_8	92.146	91.2	8.8	1.285	42 750	40 820
Xylene (pure) C_8H_{10}	106.167	90.5	9.5	1.313	43 000	40 780
Benzene I on sale [a]	–	92.1	7.9	1.26	41 870	40 190
Benzene II on sale [b]	–	91.6	8.4	1.30	42 290	40 400
Naphtalene (pure) $C_{10}H_8$ (melting temp. 80 °C)	128.19	93.7	6.3	1.20	40 360	38 940
Tetralin $C_{10}H_{12}$	132.21	90.8	9.2	1.30	42 870	40 820
Pentane C_5H_{12}	72.150	83.2	16.8	1.60	49 190	45 430
Hexane C_6H_{14}	86.177	83.6	16.4	1.584	48 360	44 670
Heptane C_7H_{16}	100.103	83.9	16.1	1.571	47 980	44 380
Octane C_8H_{18}	114.230	84.1	15.9	1.562	48 150	44 590
Benzine (mean values)	–	85	15	1.53	46 050	42 700

[a] 0.84 benzene, 0.31 toluene, 0.03 xylene (mass fractions)
[b] 0.43 benzene, 0.46 toluene, 0.11 xylene (mass fractions)

Table 4.20 Combustion of some simple gases at 25 °C and 1.01325 bar

Gas	Molar mass [a]	Density	Characteristic	Calorific value (MJ/kg)	
	(kg/kmol)	(kg/m^3)	σ	Gross	Net
Hydrogen H_2	2.0158	0.082	∞	141.80	119.97
Carbon monoxide CO	28.0104	1.14	0.50	10.10	10.10
Methane CH_4	16.043	0.656	2.00	55.50	50.01
Ethane C_2H_6	30.069	1.24	1.75	51.88	47.49
Propane C_3H_8	44.09	1.80	1.67	50.35	46.35
Butane C_4H_{10}	58.123	2.37	1.625	49.55	45.72
Ethylene C_2H_4	28.054	1.15	1.50	50.28	47.15
Propylene C_3H_6	42.086	1.72	1.50	48.92	45.78
Butylene C_4H_8	56.107	2.90	1.50	48.43	45.29
Acetylene C_2H_2	26.038	1.07	1.25	49.91	48.22

[a] According to DIN 51850: gross and net calorific values of gaseous fuels, April 1980

chamber must be equal to the enthalpy of the discharged exhaust gas.

$$\Delta h_{\mathrm{ncv}}\left[c_{\mathrm{fuel}}\right]_{25\,^\circ\mathrm{C}}^{t_{\mathrm{fuel}}}(t_{\mathrm{fuel}}-25\,^\circ\mathrm{C}) + l\left[\overline{C}_{p\,\mathrm{air}}\right]_{25\,^\circ\mathrm{C}}^{t_{\mathrm{air}}}(t_{\mathrm{air}}-25^\circ\mathrm{C}) = n_{\mathrm{exh}}\left[\overline{C}_{p\,\mathrm{exh}}\right]_{25\,^\circ\mathrm{C}}^{t}(t-25\,^\circ\mathrm{C})\,. \quad (4.107)$$

This equation includes the temperatures t_{fuel} of the fuel and t_{air} of the air, the theoretical combustion temperature t, the mean specific heat $\left[c\right]_{25\,^\circ\mathrm{C}}^{t_{\mathrm{fuel}}}$ of the fuel, and the mean specific heats $\left[\overline{C}_{p\,\mathrm{air}}\right]_{25\,^\circ\mathrm{C}}^{t_{\mathrm{air}}}$ of air and $\left[\overline{C}_{p\,\mathrm{exh}}\right]_{25\,^\circ\mathrm{C}}^{t}$ of the exhaust gas. The latter consists of the mean molar specific heats of the single

components

$$n_{\text{exh}}\left[\overline{C}_{p\,\text{exh}}\right]_{25\,^\circ\text{C}}^{t} = \frac{c}{12}\left[\overline{C}_{p\text{CO}_2}\right]_{25\,^\circ\text{C}}^{t} + \left(\frac{h}{2}+\frac{w}{18}\right)\left[\overline{C}_{p\text{H}_2\text{O}}\right]_{25\,^\circ\text{C}}^{t} + \frac{s}{32}\left[\overline{C}_{p\text{SO}_2}\right]_{25\,^\circ\text{C}}^{t} + (\lambda-1)\,o_{\min}\left[\overline{C}_{p\text{O}_2}\right]_{25\,^\circ\text{C}}^{t} + 0.79l\left[\overline{C}_{p\text{N}_2}\right]_{25\,^\circ\text{C}}^{t}\,. \quad (4.108)$$

The theoretical combustion temperature must be determined iteratively from (4.107) and (4.108). The actual combustion temperature is, even with complete combustion of the fuel, lower than the theoretical combustion temperature due to heat transfer to the environment, mainly by radiation. Also lowering the combustion temperature is the break-up of molecules (dissociation) starting above 1500 °C and the considerable dissociation above 2000 °C. The dissociation heat is released again when the temperature decreases below the dissociation temperature.

4.8.2 Internal Combustion Cycles

In internal combustion cycles, the combustion gas serves as a working fluid. It does not operate through a closed process but is discharged as exhaust gas to the environment after performing work in a turbine or a piston engine. Open gas turbine cycles and internal combustion engines (Otto and Diesel), as well as fuel cells, are internal combustion cycles. The quality of the energy transformation is assessed by the total energy efficiency

$$\eta = -P/(\dot{m}_{\text{fuel}}\Delta h_{\text{ncv}})\,,$$

where P is the power output of the cycle, $\dot{m}$ is the mass flow rate of the supplied fuel, and Δh_{ncv} is its net calorific value. The total exergetic efficiency $\xi = -P/(\dot{m}_{\text{fuel}} w_{\text{ex,fuel}})$ specifies what fraction of the exergy flow coming with the fuel is transformed into power output. Generally, w_{ex} is only slightly larger than the net calorific value, and η and ξ thus hardly differ in their numerical values. The typical total efficiency is approximately 42% for large engines (Diesel), 25% for automotive engines, and 20–30% for open gas turbine cycles.

Open Gas Turbine Cycle

In an open gas turbine plant, the inlet air is brought to a high pressure through a compressor, then preheated and heated in a combustion chamber via the combustion of the injected fuel. The combustion gases pass through a turbine, in which they do work (against the blades). The gas exiting the turbine is used to preheat the combustion air in a heat exchanger, and is then discharged into the environment. The compressor and turbine are placed on the same shaft. The power output is transformed into electric energy by a generator, which is connected to the shaft.

Otto Engine

Figure 4.17 shows the cycle of an Otto engine in p–V and T–S diagrams. At the end of the intake stroke, the cylinder is filled with a combustible fresh air–fuel

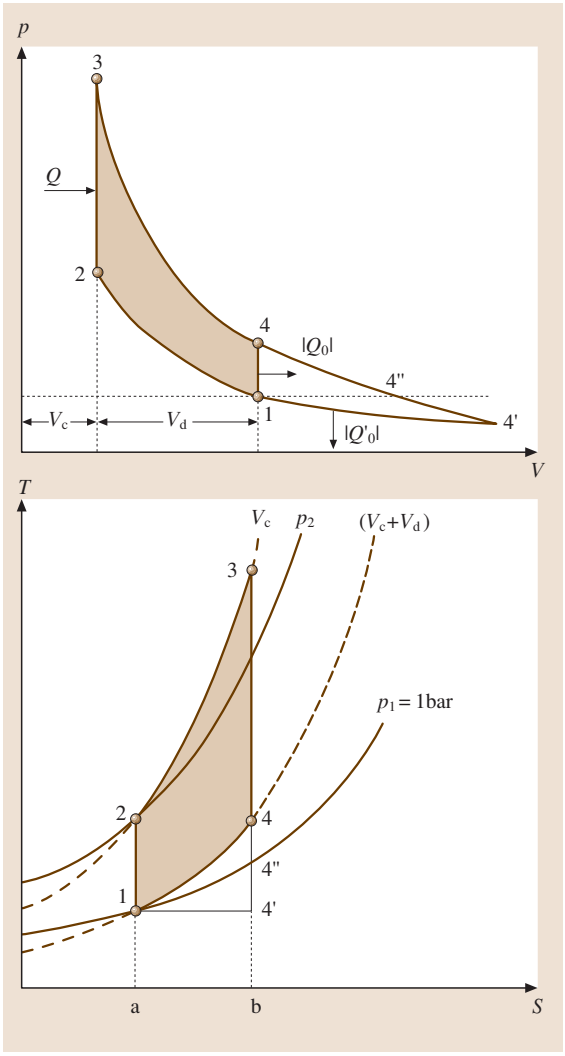


Fig. 4.17 Theoretical process of the Otto engine on p–V and the T–S diagrams

mixture of state 1 at the environment temperature and atmospheric pressure. The mixture is compressed along the adiabate 1–2 from the initial volume $V_c + V_d$ to the compression volume V_c where V_d is the displacement volume. At the top dead center 2, combustion is initiated by electric spark ignition, whereby the pressure rises from state 2 to state 3. This change of state takes place so quickly that it can be assumed to be isochoric. In Fig. 4.17 (simplifying) it is assumed that the gas is not changed and that the heat released during combustion $Q_{23} = Q$ is supplied from the outside. The gas expands along the adiabatic 3–4–4"–4' and forces the piston to return. The exhaust beginning in state 4 is substituted by the removal of energy by heat transfer $|Q_0|$ at constant volume, whereas the pressure decreases from state 4 to state 1. In state 1, the combustion gases have to be replaced by a new mixture. In order to do so, twin stroke (not shown) is necessary in a four-stroke Otto engine.

The heat transfer to the gas is

$$Q = Q_{23} = mc_v(T_3 - T_2) \tag{4.109}$$

and from the gas is

$$|Q_0| = |Q_{41}| = mc_v(T_4 - T_1)\,. \tag{4.110}$$

The work is

$$|W_t| = Q - |Q_0|\ , \tag{4.111}$$

and the thermal efficiency is given by

$$\eta = \frac{|W_t|}{Q} = 1 - \frac{T_4 - T_1}{T_3 - T_2} = 1 - \frac{T_1}{T_2}$$
$$= 1 - \left(\frac{p_1}{p_2}\right)^{\frac{\varkappa-1}{\varkappa}} = 1 - \frac{1}{\varepsilon^{\varkappa-1}}\,. \tag{4.112}$$

The compression ratio $\varepsilon = V_1/V_2 = (V_c + V_d)/V_c$ specifies the degree of adiabatic compression of the mixture. Thus, the thermal efficiency depends, except for the adiabatic exponent, only on the pressure ratio p_2/p_1 or the compression ratio ε and not on the amount of energy supplied by heat transfer. The compression ratio is limited by the self-ignition temperature of the fuel–air mixture.

Diesel Engine

The limitation to moderate compression ratios and pressures does not exist for the Diesel engine, in which the high compression heats the combustion air above the self-ignition temperature of the fuel that is injected into the hot air. Figure 4.18 shows the simplified process of the Diesel engine. It consists of the adiabatic compression 1–2 of the combustion air, isobaric combustion 2–3' after the injection of the fuel into the hot, compressed combustion air, adiabatic relaxation 3'–4, and ejection 4–1 of the exhaust gases, which is replaced in Fig. 4.18 by an isochore with heat removal. The supplied heat transfer is

$$Q_{23'} = Q = mc_p(T_{3'} - T_2) \tag{4.113}$$

and the removed heat transfer is

$$|Q_{41}| = |Q_0| = mc_v(T_4 - T_1) \tag{4.114}$$

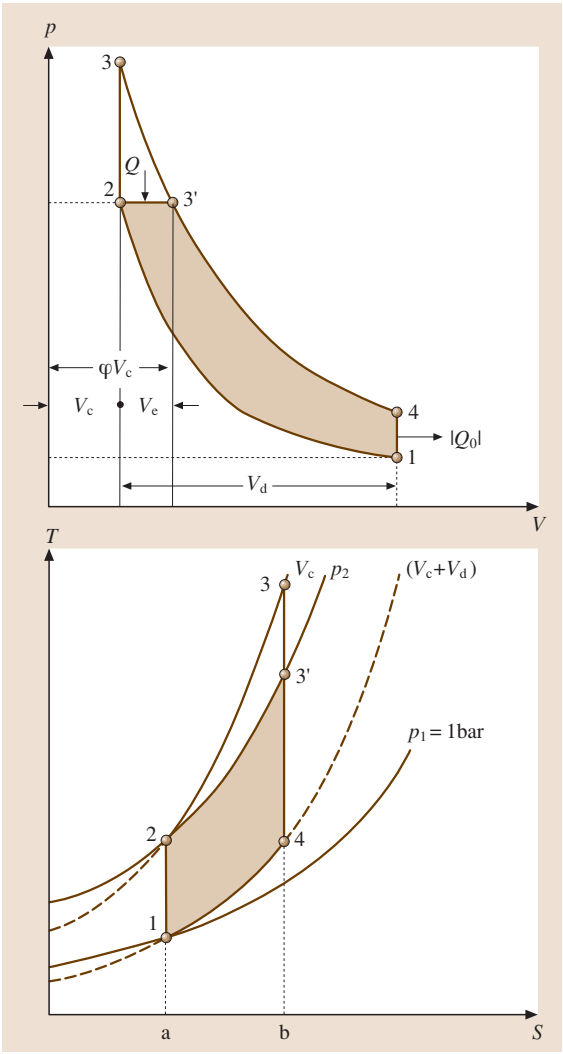


Fig. 4.18 Theoretical process of the Diesel engine on p–V and T–S diagrams

during the imaginary isochore 4–1. The work is given by

$$|W_t| = Q - |Q_0|$$

and the thermal efficiency by

$$\eta = \frac{|W_t|}{Q} = 1 - \frac{1}{\varkappa}\frac{T_4 - T_1}{T_{3'} - T_2} = 1 - \frac{1}{\varkappa}\frac{\frac{T_4}{T_3}\frac{T_3}{T_2} - \frac{T_1}{T_2}}{\frac{T_{3'}}{T_2} - 1} . \tag{4.115}$$

With the compression ratio $\varepsilon = V_1/V_2 = (V_c + V_d)/V_c$ and the cutoff ratio $\varphi = (V_c + V_e)/V_c$, the following equation for the thermal efficiency results

$$\eta = 1 - \frac{1}{\varkappa\,\varepsilon^{\varkappa-1}}\frac{\varphi^{\varkappa} - 1}{\varphi - 1} . \tag{4.116}$$

The thermal efficiency of the Diesel cycle depends, except for the adiabatic exponent, only on the compression ratio ε and on the cutoff ratio φ, which increases with increasing load.

Fuel Cells

In a fuel cell, hydrogen reacts electrochemically with oxygen to produce water

$$H_2 + \frac{1}{2}O_2 \rightarrow H_2O .$$

In this so-called cold combustion, the chemical bond energy is transformed directly into electrical energy. Figure 4.19 shows, as an example, a fuel cell with a proton conductive electrolyte, where hydrogen is supplied at the side of the anode. With the help of a catalyst, it is decomposed into two protons (H^+) and two electrons (e^-). The electrons move through a load, e.g., a motor, to the cathode. The protons move through the electrolyte to the cathode, where they, supported by a catalyst, react with the supplied oxygen, O_2, and the electrons to produce water, H_2O. There is a voltage U between the anode and cathode, and the electric current $I = F\dot{n}_{el}$ with $\dot{n}_{el} = 2\dot{n}_{H_2}$ flows. F is the Faraday constant $F = 96\,485.3$ As/mol, and $\dot{n}_{el}$ is the flow rate of electrons (SI unit mol/s). The actual terminal voltage is smaller than the reversible one because of losses due to energy dissipation in the cell. The electric power of the cell is calculated from

$$\dot{Q} + P = \dot{n}_{H_2}\Delta H^R_{H_2}$$

with the flow rate $\dot{n}_{H_2}$ of the supplied hydrogen and its molar reaction enthalpy $\Delta H^R_{H_2}$ (SI unit J/mol), which is equal to the negative molar net calorific value $\Delta H_{m\,ncv} = M_{H_2}\Delta h_{ncv}$ (Sect. 4.8.1). Analogous to the efficiency of other combustion plants, the efficiency of a fuel cell is defined as

$$\eta_{fc} = \frac{-P}{\dot{n}_{H_2}\,\Delta H_{m\,ncv}} ,$$

where the fuel cell is generally about 50% efficient.

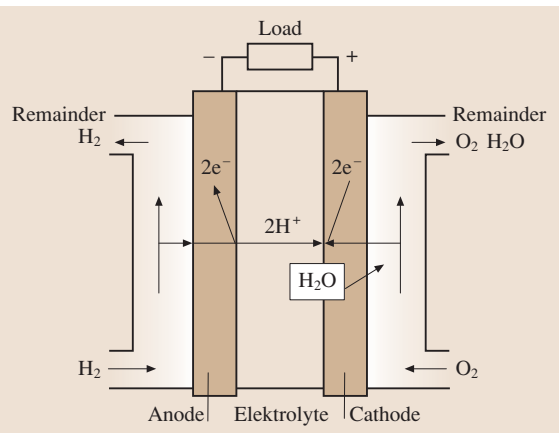


Fig. 4.19 Scheme of a fuel cell with a proton conductive electrolyte

4.8.3 Cyclic Processes, Principles

A process that brings a system back to its initial state is called a cyclic process. After the system has passed through such a cycle, all the properties of the system such as pressure, temperature, volume, internal energy, and enthalpy return to their initial values and thus produce

$$\sum Q_{ik} + \sum W_{ik} = 0 . \tag{4.117}$$

The total work done is $-W = -\sum W_{ik} = \sum Q_{ik}$.

Machines in which a fluid is undergoing a cycling process serve to transform heat transfer into work or to transfer thermal energy from a low- to a high-temperature level while work is supplied. According to the second law of thermodynamics, it is not possible to transform all the supplied heat transfer into work. If the amount of heat supplied is larger than the amount of heat discharged, the process works as a power cycle or a thermal power plant whose purpose is to deliver work. If the amount of heat discharged is smaller than the amount of heat supplied, work must be supplied. Such a process can be used for heat transfer from a medium at a lower temperature to a medium at a higher temperature, e.g., ambient temperature. The required work is

also discharged as heat at the higher temperature. Such a process works as a refrigeration cycle. In a heat pump process, heat is absorbed from the environment and is discharged together with the supplied work at a higher temperature.

Carnot Cycle

The cycle process introduced in 1824 by Carnot is shown in Figs. 4.20 and 4.21. Even though not very important in practice, the Carnot cycle played a decisive roll in the historical development of heat transfer theory. It consists of the following changes of state (here, the clockwise process of a power cycle):

1 – 2 Isothermal expansion at temperature T with heat addition Q
2 – 3 Reversible adiabatic expansion from pressure p_2 to pressure p_3
3 – 4 Isothermal compression at temperature T_0 with heat removal $|Q_0|$
4 – 1 Reversible adiabatic compression from pressure p_4 to pressure p_1

The heat supplied is

$$Q = mRT \ln V_2/V_1 = T(S_2 - S_1) \qquad (4.118)$$

and the heat removed is

$$|Q_0| = mRT_0 \ln V_3/V_4 = T_0(S_3 - S_4) = T_0(S_2 - S_1)\,. \qquad (4.119)$$

The technical work done is $-W_t = Q - |Q_0|$, and the thermal efficiency is

$$\eta = |W_t|/Q = 1 - (T_0/T)\ . \qquad (4.120)$$

With the inverse sequence 4 – 3 – 2 – 1 of changes of state, the heat absorbed Q_0 is from a body at a lower temperature and, with the supply of the technical work W_t, the heat discharged Q is at the higher temperature T. Such a counterclockwise Carnot cycle results in heat removal Q_0 from a chilled system at the low temperature T_0, thus working as a refrigerator, and can discharge the heat $|Q| = W_t + Q_0$ at the higher temperature T to the environment. If the purpose of the process is the heat release $|Q|$ at the higher temperature T for heating, the process works as a heat pump. The heat transfer Q_0 is then removed from the environment at the lower temperature T_0. Carnot cycles gained no practical importance, however, because their power related to the volume of a corresponding machine is very small. However, as an ideal, i. e., reversible, process the Carnot cycle is often used for comparison in order to assess other cyclic processes.

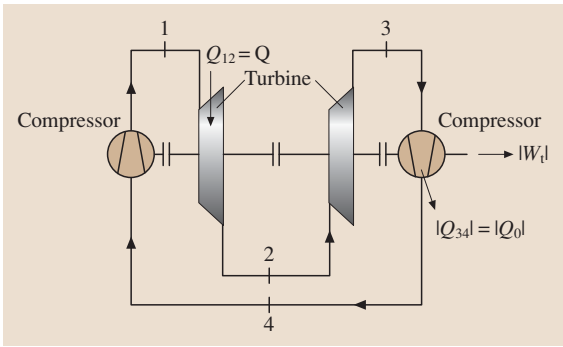


Fig. 4.20 Scheme of a Carnot power cycle

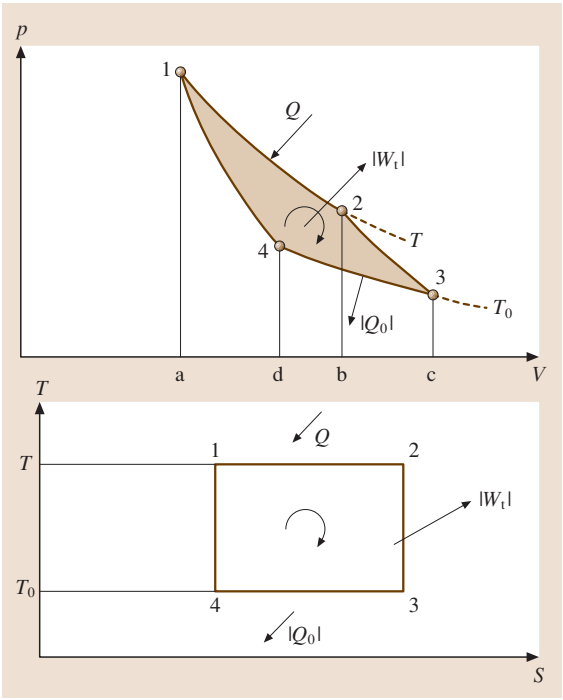


Fig. 4.21 The Carnot cycle on p–V and T–S diagrams

4.8.4 Thermal Power Cycles

In thermal power plants, energy in the form of heat transfer is transformed from the combustion gases in the working fluid, which undergoes a cyclic process. The Ackeret–Keller process consists of the following changes of state as shown in Fig. 4.22 in a p–V and T–S diagram:

1 – 2 Isothermal compression from pressure p_0 to pressure p at temperature T_0
2 – 3 Isobaric heat supply at pressure p
3 – 4 Isothermal expansion from pressure p to pressure p_0 at temperature T
4 – 1 Isobaric heat removal at pressure p_0

Because this process can be traced back to a proposal by the Swedish engineer J. Ericson (1803–1899), it is also called the Ericson cycle. It was first used in 1941, however, by Ackeret and Keller as a comparison process for gas turbine plants. The heat transfer required for the isobaric heating 2 – 3 of the compressed working fluid is provided by the isobaric cooling 4 – 1 of the expanded working fluid, $Q_{23} = |Q_{41}|$.

The thermal efficiency is equal to the efficiency of the Carnot cycle, because

$$-W_t = Q_{34} - |Q_{21}| \tag{4.121}$$

and

$$\eta = 1 - \frac{|Q_{21}|}{Q_{34}} = 1 - \frac{T_0}{T} \,. \tag{4.122}$$

However, the technical realization of this process is difficult because isothermal compression and relaxation

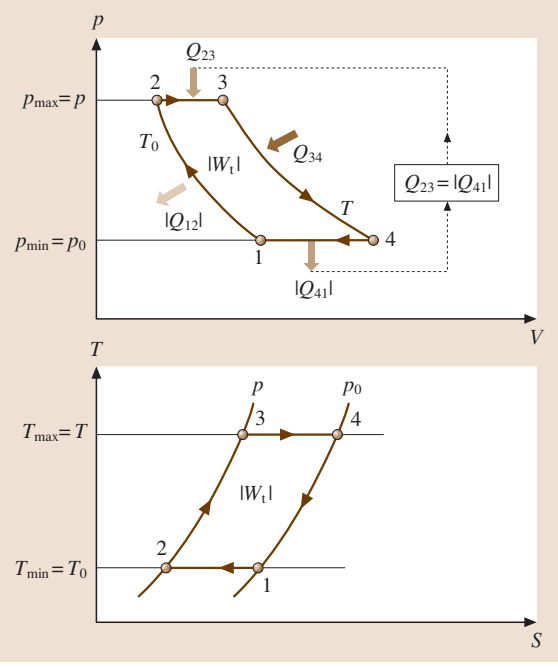


Fig. 4.22 The Ackeret–Keller process on p–V and T–S diagrams

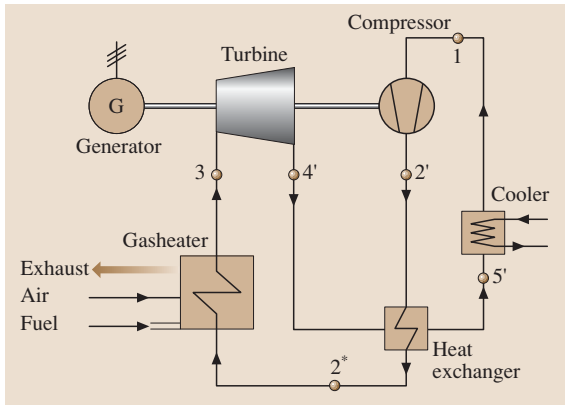


Fig. 4.23 Gas turbine process with a closed cycle

are hardly achievable due to the fact that they only can be approximated by multistage adiabatic compression with intermediate cooling. The Ackeret–Keller process serves mainly as a comparison process for the gas turbine process with multistage compression and relaxation. In a closed gas turbine plant (Fig. 4.23), a gas is compressed in the compressor, heated to a high temperature in the heat exchanger and the gas heater, then expanded in a turbine, where work is done, and cooled again to the initial temperature in the heat exchanger and in the adjacent cooler. Then the gas is drawn in by the compressor once again. Often air is used as the working fluid, but other gases such as helium or nitrogen are also sometimes used. The closed gas turbine plant is easily adjustable, and fouling of the turbine blades can be prevented by using suitable gases. A drawback in comparison to open plants is the higher energy costs, because a cooler is required and high-quality steels are needed for the heater. Figure 4.24 shows the process in the p–V and T–S diagram.

The reversible cyclic process consisting of two isobars and two isentropes is called the Joule process (states 1, 2, 3, 4). The supplied heat transfer is

$$\dot{Q} = \dot{m} c_p \, (T_3 - T_2) \,, \tag{4.123}$$

and the discharged heat transfer is

$$\left|\dot{Q}_0\right| = \dot{m} \, c_p \, (T_4 - T_1) \,. \tag{4.124}$$

The power is

$$\begin{aligned} -P &= -\dot{m} w_t = \dot{Q} - \left|\dot{Q}_0\right| \\ &= \dot{m} \, c_p (T_3 - T_2) \left(1 - \frac{T_4 - T_1}{T_3 - T_2}\right) \end{aligned} \tag{4.125}$$

and the thermal efficiency is

$$\eta = \frac{|P|}{\dot{Q}} = \left(1 - \frac{T_4 - T_1}{T_3 - T_2}\right) . \tag{4.126}$$

Because of the isentropic equation,

$$\left(\frac{p_0}{p}\right)^{\frac{\varkappa-1}{\varkappa}} = \frac{T_1}{T_2} = \frac{T_4}{T_3} \tag{4.127}$$

so

$$\frac{T_4 - T_1}{T_3 - T_2} = \frac{T_1}{T_2} = \left(\frac{p_0}{p}\right)^{\frac{\varkappa-1}{\varkappa}} \tag{4.128}$$

and the thermal efficiency is

$$\eta = \frac{|P|}{\dot{Q}} = 1 - \left(\frac{p_0}{p}\right)^{\frac{\varkappa-1}{\varkappa}} , \tag{4.129}$$

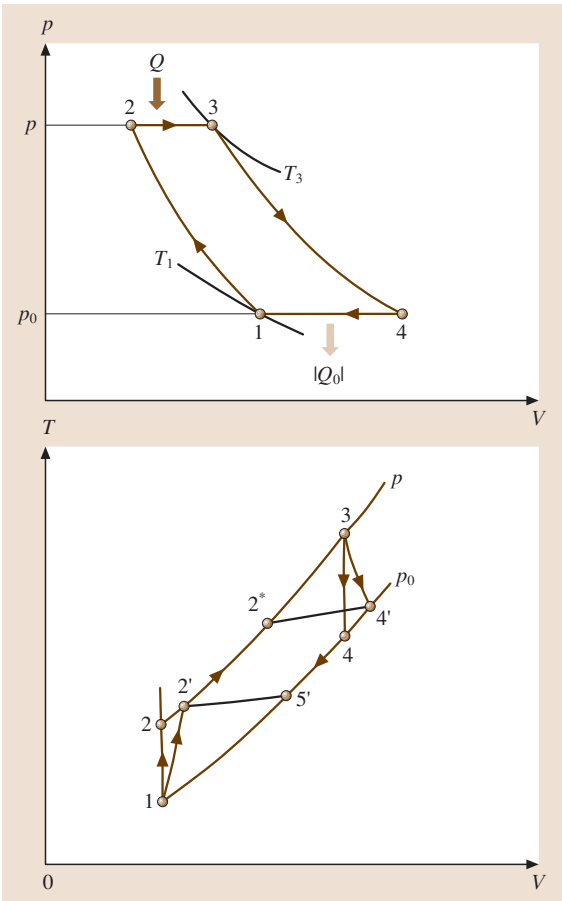


Fig. 4.24 The gas turbine process on p–V and T–S diagram. The p–V diagram shows only the reversible process (Joule process) 1, 2, 3, 4

which depends only on the pressure ratio p/p_0 or on the temperature ratio T_2/T_1 of the compression. The compressor power increases faster with the pressure ratio than does the turbine power so that the received power output according to (4.125), taking into account (4.128) becomes

$$-P = \dot{m} c_p T_1 \left[\frac{T_3}{T_1} - \left(\frac{p}{p_0}\right)^{\frac{\varkappa-1}{\varkappa}}\right]\left[1 - \left(\frac{p_0}{p}\right)^{\frac{\varkappa-1}{\varkappa}}\right] \tag{4.130}$$

which has a maximum at a certain pressure ratio for given values of the highest temperature T_3 and the lowest temperature T_1. This optimal pressure ratio follows from (4.130) through differentiation as

$$\left(\frac{p}{p_0}\right)^{\frac{\varkappa-1}{\varkappa}}_{\text{opt}} = \sqrt{\frac{T_3}{T_1}} , \tag{4.131}$$

which is, because of (4.128), equivalent to $T_4 = T_2$. Considering the efficiencies η_T of the turbine and η_C of the compressor, and the mechanical efficiency η_m for the energy transformation between turbine and compressor, the optimal pressure ratio results to

$$\left(\frac{p}{p_0}\right)^{\frac{\varkappa-1}{\varkappa}}_{\text{opt}} = \sqrt{\eta_m \eta_T \eta_C (T_3/T_1)} . \tag{4.132}$$

More than half of the turbine power of a gas turbine plant is required to drive the compressor. The completely installed power is thus four to six times the power output.

The working fluid of vapor power plants, usually water, evaporates and condenses during the process. Most electric energy is generated with such plants. The simplest form of the cyclic process (Fig. 4.25) is as follows.

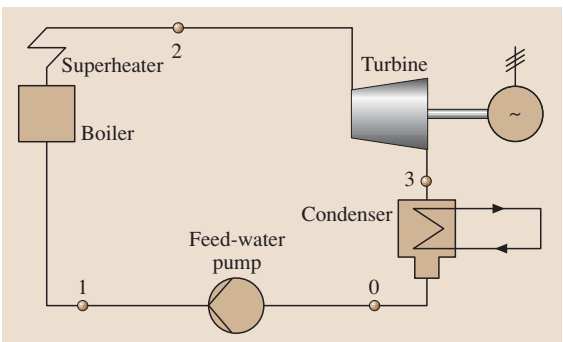


Fig. 4.25 Vapor power plant

In the boiler the working fluid is heated isobarically at a high pressure to the boiling point, evaporated, then superheated in the superheater. The vapor is then expanded adiabatically in the turbine where work is done, and condensed with heat removal in the condenser. The liquid is pressurized in the feed-water pump to the pressure of the boiler and again discharged into the boiler. The reversible cyclic process $0-1'-2-3'-0$ (Fig. 4.26), consisting of two isobars and two isentropes, is called the Clausius–Rankine process. The real cycle consists of the changes of state $0-1-2-3-0$ in Fig. 4.26. The heat absorption in the steam generator is

$$\dot{Q}_{\text{in}} = \dot{m}(h_2 - h_1) \tag{4.133}$$

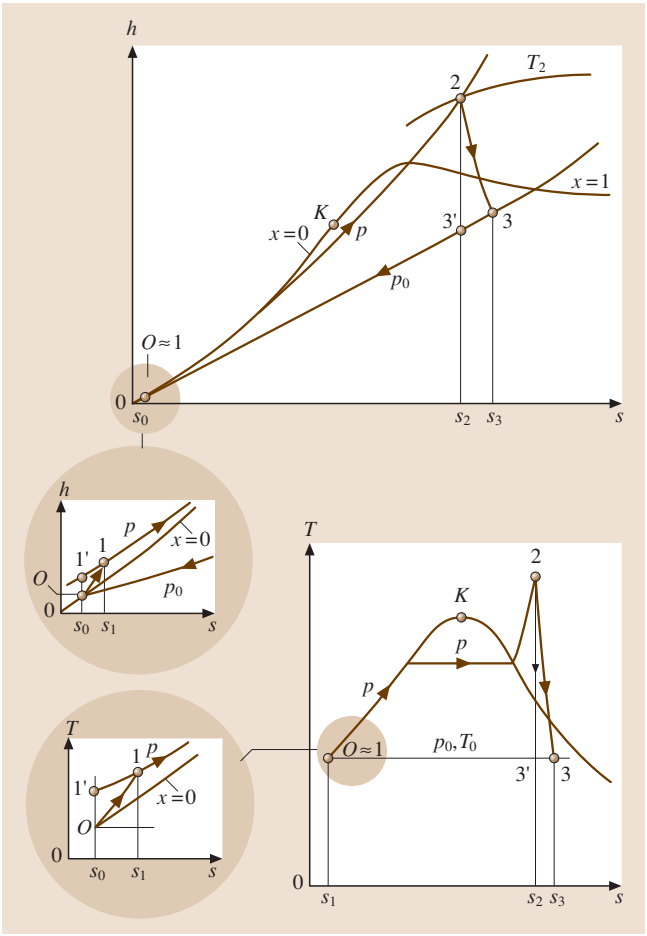


Fig. 4.26 Changes of state of the water in the cycle of a basic vapor power plant on T–S and h–s diagrams

and the power of the adiabatic turbine is

$$|P_{\text{T}}| = |\dot{m} w_{\text{t23}}| = \dot{m}(h_2 - h_3) = \dot{m}\eta_{\text{T}}(h_2 - h_3') \tag{4.134}$$

with the isentropic turbine efficiency η_{T}. The heat transfer discharged in the condenser is

$$-\dot{Q}_{\text{out}} = \dot{m}(h_3 - h_0)\,. \tag{4.135}$$

The power output of the cyclic process is

$$-P = -\dot{m} w_{\text{t}} = -P_{\text{T}} - P_{\text{P}} \tag{4.136}$$

with the pump power

$$P_{\text{P}} = \dot{m}(h_1 - h_0) = \dot{m}\frac{1}{\eta_{\text{C}}}(h_{1'} - h_0)\,, \tag{4.137}$$

where η_{C} is the efficiency of the feed-water pump. The power output differs only slightly from the power output of the turbine. The thermal efficiency is

$$\eta = -\frac{\dot{m} w_{\text{t}}}{\dot{Q}_{\text{in}}} = \frac{(h_2 - h_3) - (h_1 - h_0)}{h_2 - h_1}\,. \tag{4.138}$$

At a counter-pressure of $p_0 = 0.05$ bar, a main steam pressure of 150 bar, and a vapor temperature of 500 °C, the thermal efficiency achieves values of $\eta \approx 0.42$. Considerably higher thermal efficiencies of (presently) up to $\eta \approx 0.58$ can be achieved in combined gas–vapor power plants, in which the combustion gas at first does work in a gas turbine, where it is expanded, then is supplied to a vapor power plant in order to generate steam.

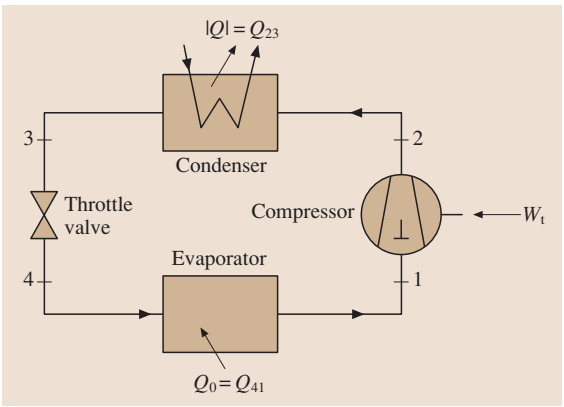


Fig. 4.27 Scheme of a vapor refrigeration plant (see text for explanation)

4.8.5 Refrigeration Cycles and Heat Pumps

Compression Refrigeration Cycle

In refrigerating machines, as well as in power plants, gases or vapors are used as working fluids. These gases/vapors are called refrigerants. A refrigeration machine is used to remove heat from a chilled system. For this purpose, it is necessary to do work, which is then transferred as heat together with the heat removed from the chilled system to the environment. For cooling with temperatures to about $-100\,^{\circ}\mathrm{C}$, compression refrigeration machines are primarily used.

Figure 4.27 shows a schematic diagram of a compression refrigeration machine. The compressor which is usually a piston compressor for small powers and a turbo compressor for large powers, draws in vapor from the evaporator at the pressure p_0 and the corresponding saturation temperature T_0 and compresses it along adiabate $1-2$ (Fig. 4.28) to pressure p. The vapor is then liquefied at pressure p in the condenser. The liquid refrigerant is expanded in the throttle valve and returns to the evaporator, where it is supplied with heat. The refrigeration machine removes from the chilled system the heat transfer q_0, which is transferred to the evaporator. In the condenser, the heat transfer $|q| = q_0 + w_t$ is transferred to the environment.

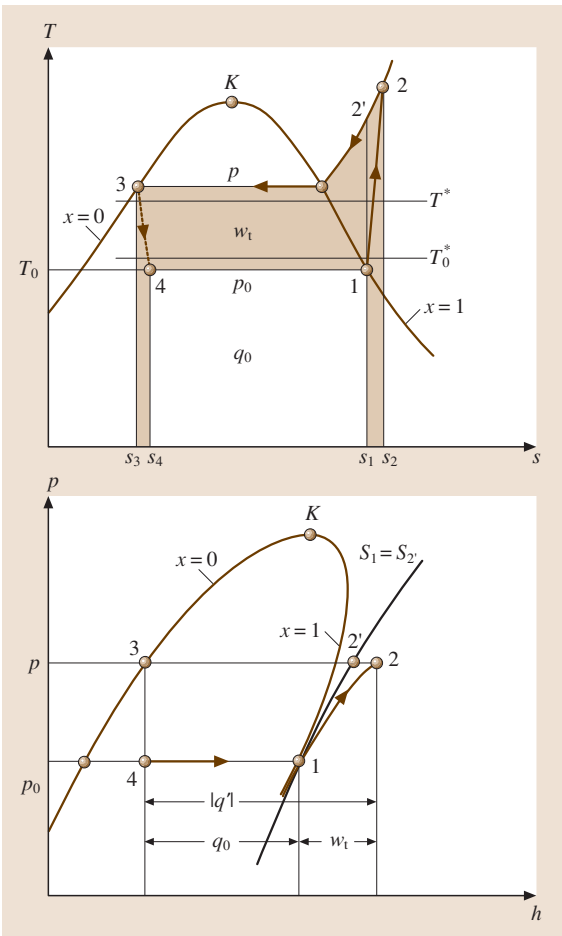


Fig. 4.28 Cycle of the refrigerant in a vapor refrigeration plant in the T–S and the p–h (Mollier) diagram

Since water freezes at $0\,^{\circ}\mathrm{C}$, and water vapor has an inconveniently large specific volume, other fluids such as ammonia NH_3, carbon dioxide CO_2, propane C_3H_8, butane C_4H_{10}, tetrafluoroethane $C_2H_2F_4$, and difluorochlormethane CHF_2Cl are used as refrigerants. Saturated refrigerant properties are given in Tables 4.10–4.13. For mass flow $\dot{m}$ of the circulating refrigerant, the refrigeration capacity is

$$\dot{q}_0 = \dot{m} q_0 = \dot{m}(h_1 - h_4) = \dot{m}\left[h''(p_0) - h'(p)\right] , \tag{4.139}$$

since $h_4 = h_3 = h'(p)$. The required power for the compressor is

$$\begin{aligned} P_C &= \dot{m} w_{t12} = \dot{m}(h_2 - h_1) \\ &= \dot{m}\frac{1}{\eta_C}\left[h_{2'} - h''(p_0)\right] , \end{aligned} \tag{4.140}$$

where η_C is the isentropic efficiency of the compressor. The heat transfer from the condenser is given by

$$|\dot{q}| = \dot{m}\,|q| = \dot{m}(h_2 - h_3) = \dot{m}\left[h_2 - h'(p)\right] . \tag{4.141}$$

The coefficient of performance of a refrigeration machine is defined as the ratio of the refrigeration capacity $\dot{q}_0$ to the required power P of the compressor

$$\varepsilon_R = \frac{\dot{q}_0}{P_C} = \frac{q_0}{w_{t12}} = \eta_C \frac{h''(p_0) - h'(p)}{h_{2'} - h''(p_0)} , \tag{4.142}$$

which depends, besides on the isentropic compressor efficiency, only on the two pressures p and p_0.

Compression Heat Pump

A compression heat pump works according to the same process as the compression refrigeration system shown in Figs. 4.27 and 4.28, where its purpose, however, is for heating. In order to provide heating, then, the heat transfer q_0 (anergy) is from the environment and is, together with the done work, w_t (the exergy), supplied as

a heat transfer $|q| = q_0 + w_t$ to the heated system. The coefficient of performance of a heat pump is defined as the ratio of the heating output $|\dot{q}|$ to the required power P of the compressor

$$\varepsilon_{hp} = \frac{|\dot{q}|}{P} = \frac{|q|}{w_t} = \eta_V \frac{h_2 - h'(p)}{h_{2'} - h''(p_0)} \,. \qquad (4.143)$$

As shown in the T–S diagram in Fig. 4.28, the area representing w_t becomes smaller at a high ambient temperature T_0^* and at a low heating temperature T^* because less power is required for the compressor and the coefficient of performance increases. In order to run heat pumps economically for the heating of housing spaces, the heating temperature must be kept low, e.g., with a floor heating at $t^* \lessapprox 29\,°C$. Additionally, heat pumps become uneconomic when the environment temperature is too low. If the coefficient of performance decreases below about 2.3, no primary energy is saved when compared to conventional heating, because the mean efficiencies for the transformation of primary energy P_{Pr} into electrical energy P in a power plant in order to run the heat pump $\eta_{el} = P/P_{Pr}$ are typically about 0.4. In that case, the heating coefficient $\xi = |\dot{q}|/P_{Pr}$ of 0.92 corresponds to the efficiency of conventional heating. Today's electrically driven heat pumps rarely achieve heating coefficients of 2.3 in the annual mean, unless the heat pump is switched off at ambient temperatures lower than approximately 3 °C and the housing space is heated conventionally. Motor-driven heat pumps with waste heat recovery and sorption heat pumps exploit the primary energy better than electrically driven heat pumps.

4.8.6 Combined Power and Heat Generation (Co-Generation)

The generation of thermal energy and electrical energy in heating power plants is called combined power and heat generation. A large amount of a power plant's waste heat, which arises in the process, is used for heating. Since the heat required for heating consists mainly – more than 90% – of anergy, less primary energy, which consists mainly of exergy, is transformed into thermal energy than in conventional heating. Low-pressure vapor is discharged from the vapor turbine; it contains, in addition to the anergy, so much exergy that the heating energy and the exergy losses in the heat distribution – normally in a long-distance heating network – are covered. Even though, compared with a simple power plant, operation work is lost due to the vapor withdrawal, the primary energy consumption for the simultaneous generation of work and thermal energy is smaller than the separate generation of work in a power plant and of thermal energy in a conventional heating system. A simplified circuit is shown in Fig. 4.29. Depending on the kind of circuit used, heating coefficients, $\xi = |\dot{q}|/P_{Pr}$, up to about 2.2 are accessible [4.20], whereas P_{Pr} is only the fraction of the primary energy that accounts for the heating. The heating coefficients are considerably above those of most heating pump systems.

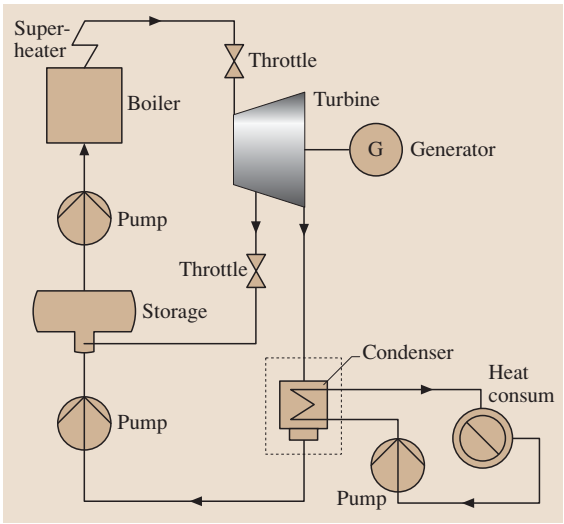


Fig. 4.29 Scheme of combined power and heat generation in extraction back-pressure operation

4.9 Ideal Gas Mixtures

A mixture of ideal gases that do not react chemically with each other also behaves as an ideal gas. The following equation of state holds

$$pV = nR_u T \,. \tag{4.144}$$

Each single gas, called a component, spreads over the entire space V as though the other gases were not present. Thus, the following equation holds for each component

$$p_i V = n_i R_u T \,, \tag{4.145}$$

where p_i is the pressure exerted by each gas individually, which is referred to as the partial pressure. The sum of all the *partial pressures* leads to $\sum p_i V = \sum n_i R_u T$ or $V \sum p_i = R_u T \sum n_i$. Comparison with (4.144) shows that

$$p = \sum p_i \tag{4.146}$$

holds. In other words, the total pressure p of the gas mixture is equal to the sum of the partial pressures of the single gases, if each gas occupies the volume V of the mixture at temperature T (Dalton's law). The thermal equation of state of an ideal gas mixture can also be written as

$$pV = mRT \,, \tag{4.147}$$

with the gas constant R of the mixture

$$R = \sum R_i m_i / m \,. \tag{4.148}$$

Specific (related to the mass in kg) caloric properties of a mixture at pressure p and temperature T result from adding the caloric properties at the same values p, T of the single gases according to their mass fractions, or

$$c_v = \frac{1}{m}\sum m_i c_{vi} \,, \qquad c_p = \frac{1}{m}\sum m_i c_{pi} \,,$$
$$u = \frac{1}{m}\sum m_i u_i \,, \qquad h = \frac{1}{m}\sum m_i h_i \,. \tag{4.149}$$

An exception to this general rule is entropy. During the mixing of single gases of state p, T to a mixture of the same state, an entropy increase takes place. This process is described by the following relation

$$s = \frac{1}{m}\left(\sum m_i s_i - \sum m_i R_i \ln \frac{n_i}{n}\right) \,, \tag{4.150}$$

where n_i is the number of moles of the single gases and n is the number of moles of the mixture. Consequently, $n_i = m_i/M_i$ and $n = \sum n_i$ with the mass m_i and the molar mass M_i of the single gases. Mixtures of real gases and liquids deviate from the above relations, in particular at higher pressures.

4.9.1 Mixtures of Gas and Vapor. Humid Air

Mixtures of gases and easily condensable vapors occur often in physics and in technology. Atmospheric air consists mostly of dry air and water vapor. Drying and climatization processes are governed by the laws of vapor–air mixtures. This holds true in the same way for the formation of fuel and vapor–air mixtures in a combustion engine. The following is limited to the examination of atmospheric air. Dry air consists of 78.04 mol% nitrogen, 21.00 mol% oxygen, 0.93 mol% argon, and 0.03 mol% carbon dioxide. Atmospheric air can be considered as a binary mixture of dry air and water, which can be present as vapor, liquid, or solid. This mixture is also called humid air. Dry air is considered a uniform substance. Since the total pressure during changes of state is almost always close to atmospheric pressure, it is possible to consider humid air, consisting of dry air and water vapor, as a mixture of ideal gases. The following relation then holds for dry air and water vapor

$$p_{air} V = m_{air} R_{air} T \quad \text{and} \quad p_v V = m_v R_v T \,. \tag{4.151}$$

These equations, together with $p = p_{air} + p_v$, allows for the determination of the mass of water vapor which is added to 1 kg dry air.

$$x_v = \frac{m_v}{m_{air}} = \frac{R_{air} p_v}{R_v (p - p_v)} \,. \tag{4.152}$$

The quantity $x_v = m_v/m_{air}$ is called the absolute or specific humidity. This quantity must not be confused with the quality x for mixtures of vapors and liquid. If water in the air is not only present as vapor, but also as liquid or solid, the water content x must be distinguished from the specific humidity x_v. The water content is defined as

$$x = \frac{m_w}{m_{air}} = \frac{m_v + m_\ell + m_{ice}}{m_{air}} = s_v + x_\ell + x_{ice} \,, \tag{4.153}$$

where m_v denotes the vapor mass, m_ℓ, the liquid mass, and m_{ice}, the ice mass in the dry air of mass m_{air}. The value x_v is the specific humidity (vapor content), x_ℓ, the liquid content, and x_{ice}, the ice content. The water content can lie between 0 (dry air) and ∞ (pure water). If

Table 4.21 Partial pressure p_{vs}, specific humidity x_s, and enthalpy h_{1+x} of saturated humid air of temperature t related to 1 kg dry air at a total pressure of 1000 mbar

t (°C)	p_{vs} (mbar)	x_s (g/kg)	h_{1+x} (kJ/kg)	t (°C)	p_{vs} (mbar)	x_s (g/kg)	h_{1+x} (kJ/kg)
−20	1.032	0.64290	−18.5164	21	24.877	15.876	61.4240
−19	1.136	0.70776	−17.3503	22	26.447	16.906	65.0741
−18	1.249	0.77825	−16.1700	23	28.104	17.995	68.8823
−17	1.372	0.85499	−14.9741	24	29.850	19.148	72.8537
−16	1.506	0.93862	−13.7609	25	31.691	20.367	77.0006
−15	1.652	1.02977	−12.5288	26	33.629	21.656	81.3286
−14	1.811	1.12906	−11.2762	27	35.670	23.019	85.8505
−13	1.984	1.23713	−10.0015	28	37.818	24.460	90.5757
−12	2.172	1.35462	−8.7030	29	40.078	25.983	95.5160
−11	2.377	1.48277	−7.3777	30	42.455	27.592	100.683
−10	2.598	1.62099	−6.0269	31	44.953	29.292	106.088
−9	2.838	1.77117	−4.6459	32	47.578	31.088	111.745
−8	3.099	1.93456	−3.2314	33	50.335	32.985	117.668
−7	3.381	2.11120	−1.7834	34	53.229	34.988	123.869
−6	3.686	2.30235	−0.2987	35	56.267	37.104	130.368
−5	4.017	2.50993	1.2277	36	59.454	39.338	137.179
−4	4.374	2.73398	2.7960	37	62.795	41.697	144.317
−3	4.760	2.97640	4.4109	38	66.298	44.188	151.805
−2	5.177	3.23851	6.0758	39	69.969	46.819	159.662
−1	5.626	3.52097	7.7926	40	73.814	49.597	167.907
0	6.117	3.8303	9.5778	41	77.840	52.530	176.563
1	6.572	4.1167	11.3064	42	82.054	55.628	185.654
2	7.061	4.4251	13.0915	43	86.464	58.901	195.208
3	7.581	4.7540	14.9290	44	91.076	62.358	205.248
4	8.136	5.1046	16.8222	45	95.898	66.009	215.806
5	8.726	5.4781	18.7741	46	100.94	69.868	226.912
6	9.354	5.8759	20.7884	47	106.21	73.947	238.603
7	10.021	6.2993	22.8684	48	111.71	78.259	250.913
8	10.730	6.7497	25.0181	49	117.45	82.817	263.878
9	11.483	7.2288	27.2416	50	123.44	87.637	277.536
10	12.281	7.7377	29.5421	51	129.70	92.743	291.958
11	13.129	8.2791	31.9263	52	136.23	98.149	307.175
12	14.027	8.8534	34.3956	53	143.03	103.87	323.221
13	14.979	9.4635	36.9572	54	150.12	109.92	340.176
14	15.988	10.111	39.6166	55	157.52	116.36	358.126
15	17.056	10.798	42.3778	56	165.22	123.17	377.094
16	18.185	11.526	45.2449	57	173.24	130.40	397.178
17	19.380	12.299	48.2272	58	181.59	138.08	418.457
18	20.644	13.118	51.3306	59	190.28	146.24	441.020
19	21.979	13.985	54.5595	60	199.32	154.92	464.964
20	23.388	14.903	57.9202				

humid air of temperature T is saturated with water vapor, the partial pressure of the water vapor is equal to the saturation pressure $p = p_{vs}$ at temperature T, and the specific humidity becomes

$$x_s = \frac{R_{air} p_{vs}}{R_v (p - p_{vs})} . \qquad (4.154)$$

Table 4.21 (cont.)

t (°C)	p_{vs} (mbar)	x_s (g/kg)	h_{1+x} (kJ/kg)	t (°C)	p_{vs} (mbar)	x_s (g/kg)	h_{1+x} (kJ/kg)
61	208.73	164.16	490.418	81	493.24	605.71	1687.252
62	218.51	174.00	517.474	82	513.42	656.65	1824.503
63	228.68	184.50	546.288	83	534.28	713.93	1978.817
64	239.25	195.71	577.001	84	555.85	778.83	2153.558
65	250.22	207.68	609.745	85	578.15	852.89	2352.928
66	261.63	220.51	644.782	86	601.19	938.12	2582.259
67	273.47	234.24	682.254	87	624.99	1037.15	2848.667
68	285.76	248.98	722.413	88	649.58	1153.60	3161.844
69	298.52	264.83	765.546	89	674.96	1292.27	3534.691
70	311.76	281.90	811.941	90	701.17	1460.20	3986.110
71	325.49	300.30	861.924	91	728.23	1667.55	4543.419
72	339.72	320.19	915.870	92	756.14	1929.63	5247.698
73	358.00	347.02	988.219	93	784.95	2271.51	6166.305
74	369.78	365.14	1037.670	94	814.65	2735.21	7412.089
75	385.63	390.62	1106.609	95	845.29	3400.16	9198.391
76	402.05	418.43	1181.826	96	876.88	4432.25	11 970.735
77	419.05	448.89	1264.123	97	909.45	6250.33	16 854.112
78	436.65	482.36	1354.501	98	943.01	10 297.46	27 724.303
79	454.87	519.28	1454.151	99	977.59	27 147.34	72 980.326
80	473.73	560.19	1564.509	100	1013.20	–	–

Example 4.13: What is the specific humidity of saturated humid air at a temperature of 20 °C and a total pressure of 1000 mbar?

The gas constants are $R_{air} = 0.2872\,\mathrm{kJ/kg\,K}$ and $R_v = 0.4615\,\mathrm{kJ/kg\,K}$. The saturated water temperature (Table 4.6) includes the vapor pressure, which is $p_{vs}(20\,°C) = 23.39\,\mathrm{mbar}$. It follows, then

$$x_s = \frac{0.2872 \times 23.39}{0.4615\,(1000 - 23.39)} \times 10^3\,\frac{\mathrm{g}}{\mathrm{kg}} = 14.905\,\frac{\mathrm{g}}{\mathrm{kg}}\,.$$

Other values of x_s are given in Table 4.21.

Degree of Saturation and Relative Humidity.

The degree of saturation is defined as $\Psi = x_v/x_s$, which is a relative measure of the vapor content. In meteorology, however, the relative humidity $\varphi = p_v(t)/p_{vs}(t)$ is often used. Close to saturation, the two values differ only slightly because

$$\frac{x_v}{x_s} = \frac{p_v(p - p_{vs})}{p_{vs}(p - p_v)} \quad \text{or} \quad \Psi = \varphi\frac{(p - p_{vs})}{(p - p_v)}\,.$$

At saturation, $\Psi = \varphi = 1$. If the pressure of saturated humid air is increased or if the temperature is decreased, the excess water vapor condenses. The condensed vapor drops out as fog or precipitation (rain); at temperatures below 0 °C, ice crystals (snow) arise. In this case, the water content is larger than the vapor content: $x > x_v = x_s$. The relative humidity can be determined with directly displaying instruments (e.g., a hair hygrometer) or with the help of an aspiration psychrometer.

Enthalpy of Humid Air

Since the amount of dry air remains constant during changes of state of humid air, and only the added amount of water varies as a result of thawing or evaporation, all properties are related to 1 kg dry air. The dry air contains $x = m_w/m_{air}$ kg water from which $x_v = m_v/m_{air}$ is vaporous. For the enthalpy h_{1+x} of the unsaturated ($x = x_v < x_s$) mixture of 1 kg dry air and x kg vapor it holds that

$$h_{1+x} = c_{p\,air}t + x_v(c_{p\,v}t + \Delta h_v)\,, \qquad (4.155)$$

with the constant-pressure specific heats $c_{p\,air} = 1.005\,\mathrm{kJ/kgK}$ of air and $c_{p\,v} = 1.86\,\mathrm{kJ/kgK}$ of water vapor, and the enthalpy of vaporization $\Delta h_v = 2500.5\,\mathrm{kJ/kg}$ of water at 0 °C. In the temperature range of interest between −60 °C and 100 °C, constant values of c_p can be assumed. At saturation, $x_v = x_s$ and $h_{1+x} = (h_{1+x})_s$. If the water content x is larger than the saturation content x_s at temperatures $t > 0°C$, the water

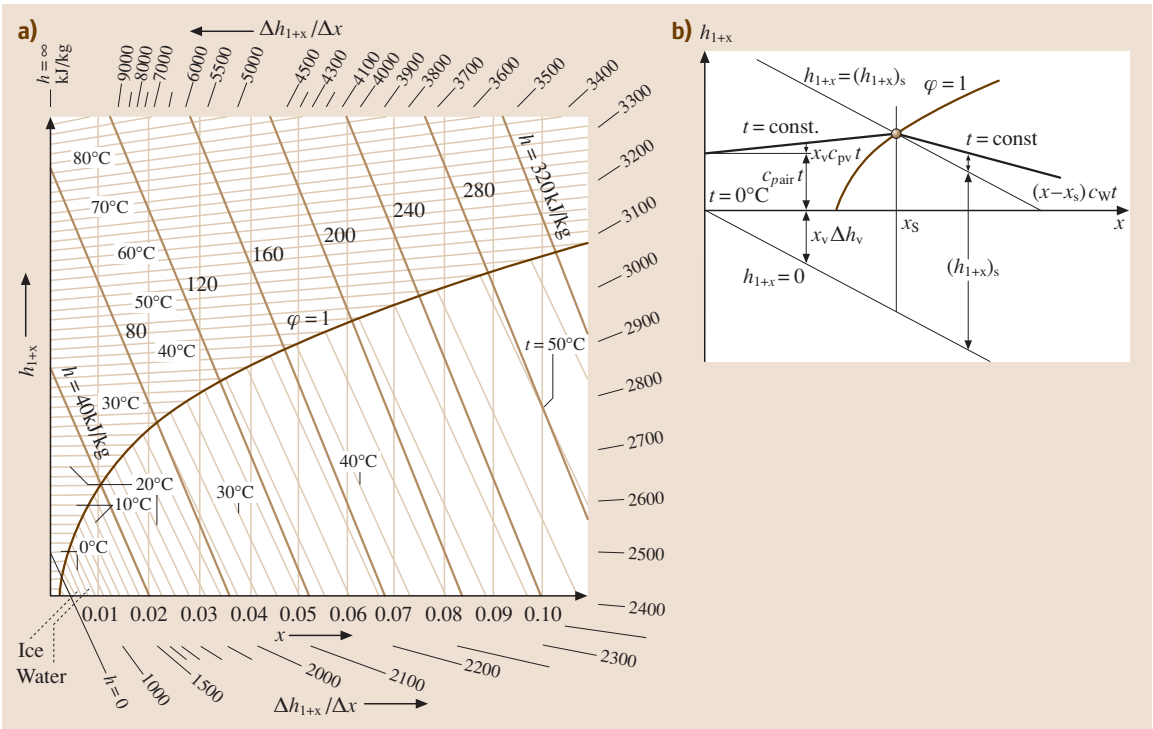


Fig. 4.30 h_{1+x}–x diagram of humid air according to Mollier

fraction $x - x_s = x_\ell$ drops out of the mixture as fog or as precipitate, and it holds that

$$h_{1+x} = (h_{1+x})_s + (x - x_s) c_w t \,. \qquad (4.156)$$

At temperatures $t < 0\,°C$, the water fraction $x - x_s = x_{ice}$ drops out as snow or ice, then

$$h_{1+x} = (h_{1+x})_s - (x - x_s)(\Delta h_f - c_{ice} t) \,. \qquad (4.157)$$

The specific heat of water is $c_w = 4.19\,kJ/kg\,K$; the specific heat of ice is $c_{ice} = 2.04\,kJ/kg\,K$; and the latent heat of fusion of ice is $\Delta h_f = 333.5\,kJ/kg$. Saturation pressures, specific humidities, and enthalpies of saturated humid air at temperatures between $-20\,°C$ and $+100\,°C$ for a total pressure of 1000 mbar are given in Table 4.21. At $t = 0\,°C$, water can be present simultaneously in all three states of aggregation. The following relation then holds for the enthalpy h_{1+x} of the mixture

$$h_{1+x} = x_s \Delta h_v - x_{ice} \Delta h_f \,. \qquad (4.158)$$

Mollier Diagram of Humid Air

Figure 4.30a shows the h_{1+x}–x diagram introduced by Mollier for the graphical depiction of changes of state of humid air. The enthalpy h_{1+x} of $(1 + x)$ kg humid air is plotted in an oblique coordinate system against the water content. The axis $h = 0$ corresponding to humid air at 0 °C is inclined right downward in such a way that the 0 °C isotherm of unsaturated humid air is horizontal. Figure 4.30b shows the construction of isotherms according to (4.155) and (4.156). Lines of constant x are vertical, while lines of constant h are straight lines parallel to the axis $h_{1+x} = 0$. Figure 4.30a includes the saturation curve $\varphi = 1$ for a total pressure of 1000 mbar. It divides the region of unsaturated mixtures (top) from the fog region (bottom), in which the humidity is contained in the mixture partly as vapor, partly as liquid (fog, precipitate) or solid (ice, fog, snow). Isotherms in the unsaturated region are, according to (4.155), towards the right, slightly ascending straight lines, which deviate at the saturation curve downward and in the fog region are nearly parallel to the straight lines of constant enthalpy, as according to (4.156). The vapor content of a state in the fog region with temperature t and water content x is determined by following isotherm t until it intersects with the saturation curve $\varphi = 1$. The fraction x_s read off in the intersection point is as vapor and thus the fraction $x - x_s$ as liquid and / or ice contained in the

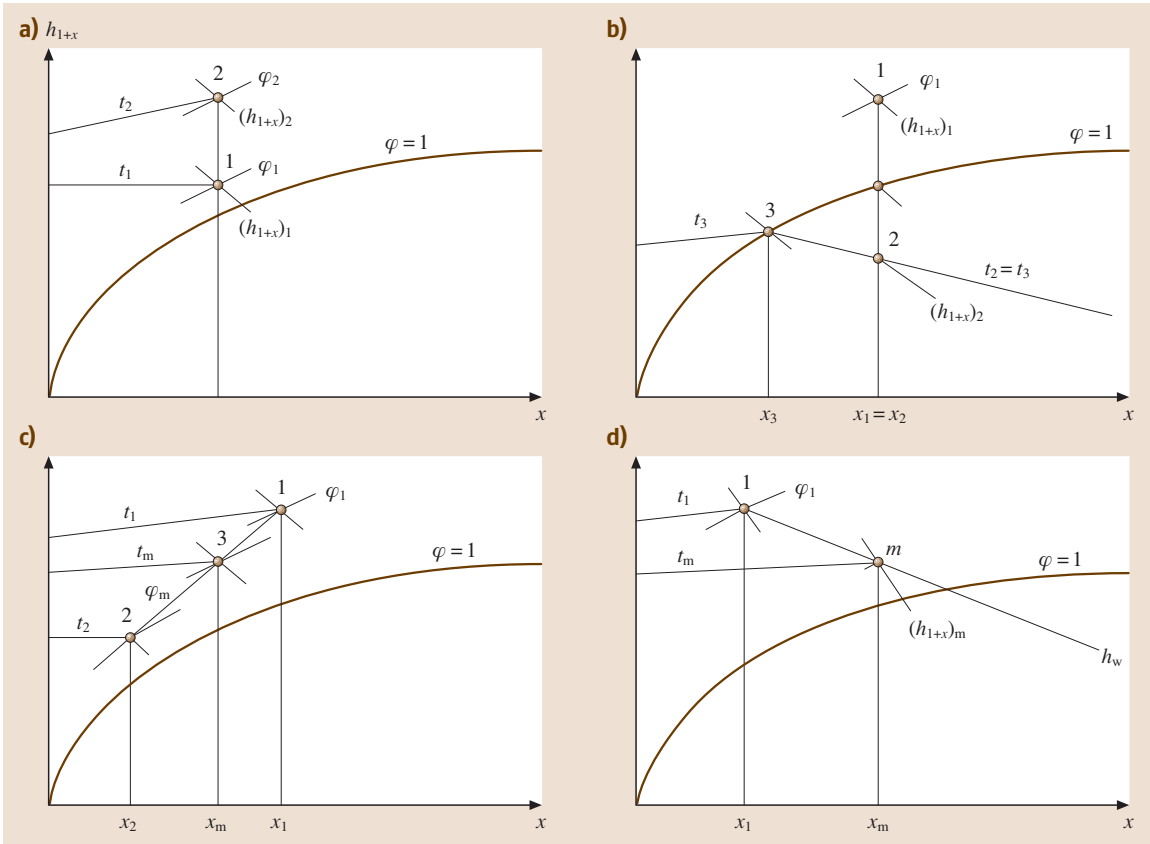


Fig. 4.31a–d Changes of state of humid air. **(a)** heating and cooling **(b)** cooling below dew point **(c)** mixture **(d)** addition of water or vapor

mixture. The inclined, beam-like pieces of straight lines $\Delta h_{1+x}/\Delta x$ determine, together with the zero-point, the direction of a change of state starting from an arbitrary state in the diagram, when water or vapor with an enthalpy in kJ/kg corresponding to the values at the boundary beams is added to the mixture. In order to find the direction of the change of state, a straight line parallel to the line determined by the origin ($h = 0,\ x = 0$) and the boundary beam must be drawn through the point of the initial state.

Changes of State of Humid Air

Heating or Cooling. If a given mixture is heated, the change of state is vertically upwards (1–2 in Fig. 4.31a). If a given mixture is cooled, the change of state is vertically downwards (2–1). As long as states 1 and 2 are in the unsaturated region, the exchanged heat related to 1 kg dry air corresponds to the vertical distance of two points of state measured in the enthalpy scale:

$$Q_{12} = m_{\text{air}}(c_{\text{p air}} + c_{\text{p v}}x)(t_2 - t_1)\,, \tag{4.159}$$

where $c_{\text{p air}} = 1.005\,\text{kJ/kg K}$ and $c_{\text{p v}} = 1.852\,\text{kJ/kg K}$. When humid air is cooled below the dew point of water (1–2 in Fig. 4.31b), precipitation drops out. The discharged heat is

$$Q_{12} = m_{\text{air}}\left[(h_{1+x})_2 - (h_{1+x})_1\right]\,, \tag{4.160}$$

where $(h_{1+x})_1$ is given by (4.155) and $(h_{1+x})_2$ by (4.156). An amount of water specified by

$$m_{\text{w}} = m_{\text{air}}(x_1 - x_3) \tag{4.161}$$

is removed.

Example 4.14: 1000 kg of humid air at $t_1 = 30\,^\circ\text{C}$, $\varphi_1 = 0.6$, and $p = 1000\,\text{mbar}$ is cooled to 15 °C. How much precipitation falls out?

The specific humidity results from (4.152) with $p_{\text{v}} = \varphi_1 p_{\text{vs}}$. According to Table 4.21, $p_{\text{vs}}\,(30\,^\circ\text{C}) =$

42.46 mbar, thus,

$$x_1 = \frac{R_{air}\,(\varphi_1\,p_{vs})}{R_v\,(p-\varphi_1\,p_{vs})} = \frac{0.2872\times0.6\times42.46}{0.4615\,(1000-0.6\times42.46)} = 16.25\times10^{-3}\,\text{kg/kg} = 16.25\,\text{g/kg}\,.$$

The 1000 kg of humid air consists of $1000/(1+x_1) = 1000/1.01625\,\text{kg} = 984.01\,\text{kg}$ dry air and $(1000-984.01)\,\text{kg} = 15.99\,\text{kg}$ water vapor. The water content at point 3, $x_3 = x_s$, follows from Table 4.21 at $t_3 = 15\,°\text{C}$ to $x_3 = 10.79\,\text{g/kg}$, thus, $m_\ell = 984.01\times(16.25-10.80)\times10^{-3}\,\text{kg} = 5.36\,\text{kg}$.

Mixture of Two Amounts of Air. If two amounts of humid air at states 1 and 2 are mixed adiabatically (i. e., without heat exchange with the environment), state m after the mixture (point 3 in Fig. 4.31c) is located on the straight line connecting states 1 and 2. Point m is determined by subdividing the straight connecting line 1–2 equivalent to the ratio of the dry air masses m_{air2}/m_{air1}. It is then

$$x_m = \frac{m_{air1}x_1 + m_{air2}x_2}{m_{air1}+m_{air2}}\,. \qquad (4.162)$$

Mixing two saturated air amounts of different temperatures always leads to the formation of fog, as the water amount $x_m - x_s$ drops out, where x_s is the specific humidity at saturation on the isotherm passing through the mixture point in the fog region.

Example 4.15: 1000 kg of humid air at $t_1 = 30\,°\text{C}$ and $\varphi_1 = 0.6$ are mixed at 1000 mbar with 1500 kg of saturated humid air at $t_2 = 10\,°\text{C}$. What is the temperature after the mixture?

As calculated in the previous example, $x_1 = 16.25\,\text{g/kg}$. The specific humidity at saturation for $t_2 = 10\,°\text{C}$ given in Table 4.21 is $x_{2s} = 7.7377\,\text{g/kg}$. The dry air masses are

$$m_{air1} = 1000/(1+x_1)\,\text{kg} = 1000/(1+16.25\times10^{-3})\,\text{kg} = 984.01\,\text{kg}\,,$$

and

$$m_{air2} = 1500/(1+x_{2s})\,\text{kg} = 1500/(1+7.7377\times10^{-3})\,\text{kg} = 1488.5\,\text{kg}\,.$$

The water content after the mixture therefore becomes

$$x_m = \frac{984.01\times16.25+1488.5\times7.7377}{984.01+1488.5}\,\text{g/kg} = 11.12\,\text{g/kg}\,.$$

The enthalpies, calculated according to (4.155), are

$$(h_{1+x})_1 = \left[1.005\times30+16.25\times10^{-3}\times(1.86\times30+2500.5)\right]\text{kJ/kg} = 71.69\,\text{kJ/kg}\,,$$

$$(h_{1+x})_2 = \left[1.005\times10+7.7377\times10^{-3}\times(1.86\times10+2500.5)\right]\text{kJ/kg} = 29.54\,\text{kJ/kg}\,.$$

The enthalpy of the mixture is

$$(h_{1+x})_m = \frac{m_{air1}(h_{1+x})_1 + m_{air2}\,(h_{1+x})_2}{m_{air1}+m_{air2}} = \frac{984.01\times71.69+1488.5\times29.54}{984.01+1488.5}\,\text{kJ/kg} = 46.31\,\text{kJ/kg}.$$

On the other hand, according to (4.155), the following also holds

$$(h_{1+x})_m = \left(1.005\,t_m + 11.12\times10^{-3}\times(1.86\,t_m+2500.5)\right)\text{kJ/kg}.$$

From this it follows that $t_m = 18\,°\text{C}$.

Addition of Water or Vapor. If humid air is mixed with m_w kg of water or water vapor, the water content after the mixture is $x_m = (m_{air1}x_1 + m_w)/m_{air1}$. The enthalpy is

$$(h_{1+x})_m = \left[m_{air\,1}(h_{1+x})_1 + m_w h_w\right]/m_{air1}\,. \qquad (4.163)$$

The final state after the mixture is located in the Mollier diagram for humid air (Fig. 4.31d) on a straight line passing through the origin with the gradient h_w, where $h_w = \Delta h_{1+x}/\Delta x$ is given by the pieces of straight lines on the boundary scale.

Wet-Bulb Temperature. When unsaturated humid air of state t_1, x_1 passes over a water or ice surface, water evaporates or ice sublimates, causing the specific humidity of the humid air to increase. During this increase in specific humidity, the temperature of the water or of the ice decreases and adopts, after a sufficiently long time, a final value, which is called the wet-bulb temperature. The wet-bulb temperature t_{wb} can be determined in the Mollier diagram by looking for the isotherm t_{wb} in the fog region whose extension passes through state 1.

4.10 Heat Transfer

If temperature differences exist between bodies that are not isolated from each other or within different areas of the same body, energy flows from the region of higher temperature to the region of lower temperature. This process is called heat transfer and will continue until the temperatures are balanced. Three modes of heat transfer are distinguished.

- Heat transfer by conduction in solids, motionless liquids, or motionless gases. Kinetic energy is hereby transferred from a molecule or an elementary particle to its neighbor.
- Heat transfer by convection in liquids or gases with bulk fluid motion.
- Heat transfer by radiation takes place in the form of electromagnetic waves and without the presence of an intervening medium.

In engineering, all three modes of heat transfer are often present at the same time.

4.10.1 Steady-State Heat Conduction

Steady-State Heat Conduction Through a Plane Wall

If different temperatures are prescribed on two surfaces of a plane wall with thickness δ, according to Fourier's law, the heat transfer

$$Q = \lambda A \frac{T_1 - T_2}{\delta} \tau$$

flows through the area A over time τ. Here, λ is a material property (SI unit W/(Km)) that is called the thermal conductivity (Table 4.22). The rate of heat transfer is given by $Q/\tau = \dot{Q}$ (SI unit W), and $Q/(\tau A) = \dot{q}$ is referred to as the heat flux (SI unit W/m^2). It holds, then

$$\dot{Q} = \lambda A \frac{T_1 - T_2}{\delta} \quad \text{and} \quad \dot{q} = \lambda \frac{T_1 - T_2}{\delta} . \tag{4.164}$$

Similar to electric conduction, where a current I flows only when a voltage U exists to overcome the resistance R ($I = U/R$), heat transfer occurs only when a temperature difference $\Delta T = T_2 - T_1$ exists

$$\dot{Q} = \frac{\lambda A}{s} \Delta T .$$

Analogous to Ohm's law, $R_{\text{th}} = \delta/(\lambda A)$ is called the thermal resistance (SI unit K/W).

Fourier's Law

Considering a layer perpendicular to the heat transfer of thickness dx instead of the wall with the finite thickness δ leads to Fourier's law in the differential form

$$\dot{Q} = -\lambda A \frac{\mathrm{d}T}{\mathrm{d}x} \quad \text{and} \quad \dot{q} = -\lambda \frac{\mathrm{d}T}{\mathrm{d}x} , \tag{4.165}$$

where the minus sign results from the fact that heat transfer occurs in the direction of decreasing temperature. Here, $\dot{Q}$ is the heat transfer in the direction of the x-axis, as is the same for $\dot{q}$. The heat flux in the direction of the three coordinates x, y, and z is given in vector form by

$$\dot{q} = -\lambda \left(\frac{\partial T}{\partial x} \boldsymbol{e}_x + \frac{\partial T}{\partial y} \boldsymbol{e}_y + \frac{\partial T}{\partial z} \boldsymbol{e}_z \right) \tag{4.166}$$

with the unit vectors $\boldsymbol{e}_x, \boldsymbol{e}_y, \boldsymbol{e}_z$. At the same time, (4.166) is the general form of Fourier's law. In this form, Fourier's law holds for isotropic materials, i. e., materials with equal thermal conductivities in the direction of the three coordinate axes.

Steady-State Heat Conduction Through a Tube Wall

According to Fourier's law, the heat transfer rate through a cylindrical area of radius r and length l is $\dot{Q} = -\lambda\, 2\pi r l (\mathrm{d}T/\mathrm{d}r)$. Under steady-state conditions, the heat transfer rate is the same for all radii and thus $\dot{Q} = \text{const}$. It is therefore possible to separate the variables T and r and to integrate from the inner surface of the cylinder, $r = r_{\text{i}}$ with $T = T_{\text{i}}$, to an arbitrary location r with temperature T. The temperature profile in a tube wall of thickness $r - r_{\text{i}}$ becomes

$$T_{\text{i}} - T = \frac{\dot{Q}}{\lambda\, 2\pi l} \ln \frac{r}{r_{\text{i}}} .$$

With temperature T_{o} at the outer surface at radius r_{o}, the heat transfer rate through a tube of thickness $r_{\text{o}} - r_{\text{i}}$ and length l becomes

$$\dot{Q} = \lambda\, 2\pi l \frac{T_{\text{i}} - T_{\text{o}}}{\ln r_{\text{o}}/r_{\text{i}}} . \tag{4.167}$$

In order to obtain formal agreement with (4.164), it is also possible to write

$$\dot{Q} = \lambda A_{\text{m}} \frac{T_{\text{i}} - T_{\text{o}}}{\delta} \tag{4.168}$$

where $\delta = r_{\text{o}} - r_{\text{i}}$ and $A_{\text{m}} = \frac{A_{\text{o}} - A_{\text{i}}}{\ln(A_{\text{o}}/A_{\text{i}})}$, if $A_{\text{o}} = 2\pi r_{\text{o}} l$ is the outer and $A_{\text{i}} = 2\pi r_{\text{i}} l$ is the inner surface of the tube.

Table 4.22 Thermal conductivities λ (W/(mK))

Solids at 20 °C	
Silver	458
Copper, pure	393
Copper, merchandized	350–370
Gold, pure	314
Aluminium (99.5%)	221
Magnesium	171
Brass	80–120
Platinum, pure	71
Nickel	58.5
Iron	67
Gray cast iron	42–63
Steel, 0.2% C	50
Steel, 0.6% C	46
Constantane, 55% Cu, 45% Ni	40
V2A, 18% Cr, 8% Ni	21
Monel metal	
67% Ni, 28% Cu, 5% Fe + Mn + Si + C	25
Manganin	22.5
Graphite, increasing with density and purity	12–175
Hard coal, natural	0.25–0.28
Stone, different kinds	1–5
Quartz glass	1.4–1.9
Concrete, Ferroconcrete	0.3–1.5
Fire resistant stones	0.5–1.7
Glass (2500) [a]	0.81
Ice, at 0 °C	2.2
Soil, clayey damp	2.33
Soil, dry	0.53
Quartz sand, dry	0.3
Brickwork, dry	0.25–0.55
Brickwork, damp	0.4–1.6
Insulating material at 20 °C	
Alfol	0.03
Asbestos	0.08
Asbestos plates	0.12–0.16
Glass wool	0.04
Cork plates (150) [a]	0.05
Diatomite, fired	0.08–0.13
Slag wool, rockwool matte (120) [a]	0.035
Slag wool, dense (?)	0.045
Synthetic resins – foams (15) [a]	0.035
Silk (100) [a]	0.055
Peat plates, air dry	0.04–0.09
Wool	0.04

[a] in brackets density in kg/m^3

Table 4.22 (cont.)

Liquids	
Water [b] of 1 bar at 0 °C	0.562
Water [b] of 1 bar at 20 °C	0.5996
Water [b] of 1 bar at 50 °C	0.6405
Water [b] of 1 bar at 80 °C	0.6668
At saturation: 99.63 °C	0.6773
Carbon dioxide at 0 °C	0.109
Carbon dioxide at 20 °C	0.086
Lubricating oils	0.12–0.18
Gases at 1 bar and temperature t in 20 °C	
Hydrogen,	
$-100\,°C \le \theta \le 1000\,°C$	$0.171(1+0.00349\theta)$
Air, $0\,°C \le \theta \le 1000\,°C$	$0.0245(1+0.00225\theta)$
Carbon dioxide,	
$0\,°C \le \theta \le 1000\,°C$	$0.01464(1+0.005\theta)$

[b] according to [4.21]

A_m is the logarithmic mean between the outer and inner tube surfaces. The thermal resistance of the tube $R_{th} = \delta/(\lambda A_m)$ (SI unit K/W) must be overcome by the temperature difference so that heat transfer occurs.

4.10.2 Heat Transfer and Heat Transmission

If heat is transferred from a fluid to a wall, conducted through the wall and, on the other side, transferred to a second fluid, this process is called heat transmission. In this case, two heat transfer processes and a heat conduction process are connected in series. There exists a steep temperature drop in a layer directly at the wall (Fig. 4.32), where the temperature changes only slightly farther away from the wall. Due to the no-slip condition

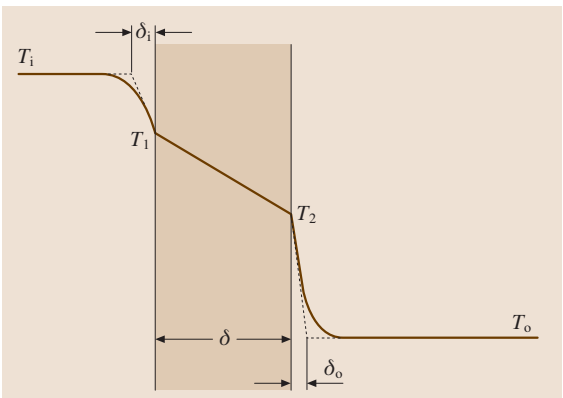


Fig. 4.32 Heat transmission through a flat wall

for the fluid at the wall surface, it can simplistically be assumed that a thin fluid boundary layer at rest, of thickness δ_i and δ_o, respectively, adheres to the wall while the fluid outside balances the temperature differences. In the thin fluid layer, heat transfer is by conduction and, according to Fourier's law, the heat flow transfer rate at the left wall side is given by

$$\dot{Q} = \lambda A \frac{T_i - T_1}{\delta_i} ,$$

where λ is the thermal conductivity of the fluid. The film thickness depends on many parameters such as the velocity of the fluid along the wall and the form and surface conditions of the wall. It has been proven suitable to use the quotient $\lambda/\delta_i = \alpha$ instead of the film thickness δ_i. This leads to the Newtonian formulation for the heat transfer rate from a fluid to a solid surface

$$\dot{Q} = \alpha A (T_f - T_0) , \qquad (4.169)$$

where T_f is the fluid temperature and T_0 is the surface temperature. The quantity α is defined as the heat transfer coefficient (SI unit $W/(m^2K)$). Orders of magnitude for heat transfer coefficients are given in Table 4.23. The basics needed for the calculation of α are contained in section Sect. 4.10.4. Following Ohm's law $I = (1/R) \times U$, the quantity $1/(\alpha A) = R_{th}$ is also called the convective heat transfer resistance (SI unit K/W). It must be overcome by the temperature difference $\Delta T = T_f - T_0$ to enable the heat transfer $\dot{Q}$. In Fig. 4.32, the heat transfer must overcome three single resistances in series, which sum up to the total resistance.

Heat Transmission Through a Plane Wall.

The heat transfer passing through a plane wall (Fig. 4.32) is given by

$$\dot{Q} = kA(T_i - T_o) , \qquad (4.170)$$

Table 4.23 Heat transfer coefficients α

Natural convection in:	α (W/m^2K)		
Gases	3	–	20
Water	100	–	600
Boiling water	1000	–	20 000
Forced convection in:			
Gases	10	–	100
Liquids	50	–	500
Water	500	–	10 000
Condensing vapor	1000	–	100 000

where $1/(kA)$ is the total heat resistance, which is, again, the sum of the individual resistances

$$\frac{1}{kA} = \frac{1}{\alpha_i A} + \frac{\delta}{\lambda A} + \frac{1}{\alpha_o A} . \qquad (4.171)$$

The quantity k, defined by (4.170) is called the heat transmission coefficient (SI unit $W/(m^2K)$). If the wall consists of several homogeneous layers (Fig. 4.33) with thicknesses δ_1, δ_2, ... and thermal conductivities λ_1, λ_2, ..., (4.170) holds likewise with the total resistance

$$\frac{1}{kA} = \frac{1}{\alpha_i A} + \sum \frac{\delta_j}{\lambda_j A} + \frac{1}{\alpha_o A} . \qquad (4.172)$$

Example 4.16: The wall of a cold store consists of a 5 cm-thick, internal concrete layer ($\lambda = 1\,W/(Km)$), a 10 cm-thick cork stone insulation ($\lambda = 0.04\,W/(Km)$), and a 50 cm-thick external brick wall. The inner heat transfer coefficient is $\alpha_i = 7\,W/(m^2K)$ and the outer coefficient is $\alpha_o = 20\,W/(m^2K)$. What is the heat transfer rate through 1 m^2 of the wall if the temperatures inside and outside are −5 °C and 25 °C, respectively?

According to (4.172) the heat transmission resistance is

$$\frac{1}{kA} = \left(\frac{1}{7\times1} + \frac{0.05}{1\times1} + \frac{0.1}{0.04\times1} + \frac{0.5}{0.75\times1} + \frac{1}{20\times1}\right)\frac{K}{W}$$
$$= 3.41\,\frac{K}{W} .$$

The heat transfer rate is $\dot{Q} = \frac{1}{3.41}(-5-25)\,W$, $|\dot{Q}| = 8.8\,W$.

Heat Transmission Through Tubes.

For heat transmission through tubes, (4.170) again holds, where the thermal resistance is the sum of the single resistances

$$\frac{1}{kA} = \frac{1}{\alpha_i A_i} + \frac{\delta}{\lambda A_m} + \frac{1}{\alpha_o A_o} .$$

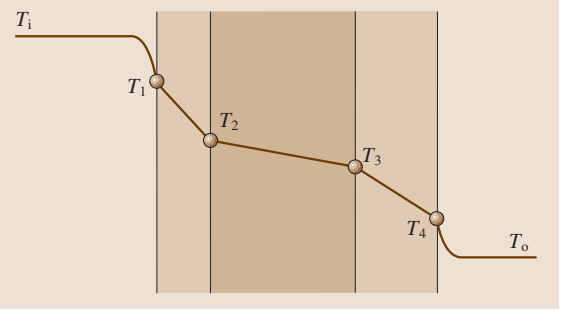


Fig. 4.33 Heat transmission through a plane, multilayered wall

Table 4.24 Material properties of liquids, gases, and solids

	t (°C)	ρ (kg/m^3)	c_p (J/kg)	λ (W/(mK))	$a \times 10^6$ (m^2/s)	$\eta \times 10^6$ (Pas)	Pr
Mercury	20	13 600	139	8000	4.2	1550	0.027
Sodium	100	927	1390	8600	67	710	0.0114
Lead	400	10 600	147	15 100	9.7	2100	0.02
Water	0	999.8	4217	0.562	0.133	1791.8	13.44
	5	1000	4202	0.572	0.136	519.6	11.16
	20	998.3	4183	0.5996	0.144	1002.6	6.99
	99.3	958.4	4215	0.6773	0.168	283.3	1.76
Thermal oil	20	887	1000	0.133	0.0833	426	576
	80	835	2100	0.128	0.073	26.7	43.9
	150	822	2160	0.126	0.071	18.08	31
Air	−20	1.3765	1006	0.02301	16.6	16.15	0.71
	0	1.2754	1006	0.02454	17.1	19.1	0.7
	20	1.1881	1007	0.02603	21.8	17.98	0.7
	100	0.9329	1012	0.03181	33.7	21.6	0.69
	200	0.7256	1026	0.03891	51.6	25.7	0.68
	300	0.6072	1046	0.04591	72.3	29.2	0.67
	400	0.5170	1069	0.05257	95.1	32.55	0.66
Water vapor	100	0.5895	2032	0.02478	20.7	12.28	1.01
	300	0.379	2011	0.04349	57.1	20.29	0.938
	500	0.6846	1158	0.05336	67.29	34.13	0.741
Aluminium 99.99%	20	2700	945	238	93.4	–	–
V2A steel, hardened and tempered	20	8000	477	15	3.93	–	–
Lead	20	11 340	131	35.3	23.8	–	–
Chrome	20	6900	457	69.1	21.9	–	–
Gold, pure	20	19 290	128	295	119	–	–
UO_2	600	11 000	313	4.18	1.21	–	–
	1000	10 960	326	3.05	0.854	–	–
	1400	10 900	339	2.3	0.622	–	–
Gravel concrete	20	2200	879	1.28	0.662	–	–
Plaster	20	1690	800	0.79	0.58	–	–
Fir, radial	20	410	2700	0.14	0.13	–	–
Cork plates	30	190	1880	0.041	0.11	–	–
Glass wool	0	200	660	0.037	0.28	–	–
Soil	20	2040	1840	0.59	0.16	–	–
Quartz	20	2300	780	1.4	0.78	–	–
Marble	20	2600	810	2.8	1.35	–	–
Chamotte	20	1850	840	0.85	0.52	–	–
Wool	20	100	1720	0.036	0.21	–	–
Hard coal	20	1350	1260	0.26	0.16	–	–
Snow (compact)	0	560	2100	0.46	0.39	–	–
Ice	0	917	2040	2.25	1.2	–	–
Sugar	0	1600	1250	0.58	0.29	–	–
Graphite	20	2250	610	155	1.14	–	–

The heat transmission coefficient k is usually related to the outer tube surface $A = A_o$, which is often easier to determine. The following equation therefore holds

$$\frac{1}{kA_o} = \frac{1}{\alpha_i A_i} + \frac{\delta}{\lambda A_m} + \frac{1}{\alpha_o A_o} , \qquad (4.173)$$

where $A_m = (A_o - A_i)/\ln(A_o/A_i)$. If the tube consists of several homogeneous layers with thicknesses $\delta_1, \delta_2, \ldots$ and thermal conductivities $\lambda_1, \lambda_2, \ldots$, (4.170) likewise holds for the total resistance

$$\frac{1}{kA_o} = \frac{1}{\alpha_i A_i} + \sum \frac{\delta_j}{\lambda_j A_{mj}} + \frac{1}{\alpha_o A_o} , \qquad (4.174)$$

where the total resistance must be summed from the single layers j with their respective mean logarithmic areas

$$A_{mj} = (A_{oj} - A_{ij}) \ln\left(\frac{A_{oj}}{A_{ij}}\right) .$$

4.10.3 Transient Heat Conduction

During transient heat conduction, the temperatures vary with respect to time. In a plane wall with prescribed surface temperatures, the temperature profile is no longer linear as the heat transfer into the wall differs from the heat transfer out. The difference between transfer in and heat transfer out increases (or decreases) the internal energy of the wall and, thus, its temperature is a function of time. For plane walls with heat transfer in the direction of the x-axis, Fourier's heat conduction equation holds

$$\frac{\partial T}{\partial \tau} = a \frac{\partial^2 T}{\partial x^2} . \qquad (4.175)$$

Multidimensional heat conduction is represented by the following relation

$$\frac{\partial T}{\partial \tau} = a \left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2} \right) . \qquad (4.176)$$

In this form, both equations assume constant thermal conductivity λ (isotropic). The quantity $a = \lambda/(\varrho\, c)$ is defined as the thermal diffusivity (SI unit m^2/s), numerical values for which are given in Table 4.24.

For the solution of Fourier's equation, it is suitable to introduce – as in other heat transfer problems – dimensionless quantities, which reduce the number of variables. Equation (4.175) is considered in order to demonstrate the basic procedure. The dimensionless temperature is set to $\Theta = (T - T_c)/(T_0 - T_c)$, where T_c is a characteristic constant temperature and T_0 is the initial temperature. If the cooling of a plate with an initial temperature T_0 in a cold environment is considered, T_c could be, for example, the ambient temperature T_{env}. All lengths are related to a characteristic length X, e.g., half of the plate thickness. Furthermore, it is suitable to introduce the dimensionless time, which is called the *Fourier number*, as $\mathrm{Fo} = a\tau/X^2$. The solution of the heat conduction equation then has the form

$$\Theta = f(x/X, \mathrm{Fo}).$$

In many problems, the heat transfer to the surface of a body by convection to the surrounding fluid of temperature T_{env}. The energy balance then holds at the surface (index w = wall)

$$-\lambda \left(\frac{\partial T}{\partial x}\right)_w = \alpha(T_w - T_{env})$$

or

$$\frac{1}{\Theta_w}\left(\frac{\partial \Theta}{\partial \xi}\right)_w = -\frac{\alpha X}{\lambda} ,$$

where $\xi = x/X$, $\Theta = (T - T_{env})/(T_0 - T_{env})$, and $\Theta_w = (T_w - T_{env})/(T_0 - T_{env})$. The solution is also a function of the dimensionless quantity $\alpha X/\lambda$, which is defined as the Biot number Bi, where the thermal conductivity λ of the body is assumed to be constant, and α is the heat transfer coefficient between the body and the surrounding fluid. Solutions of (4.175) have the form

$$\Theta = f(x/X, \mathrm{Fo}, \mathrm{Bi}) . \qquad (4.177)$$

Semi-infinite Body

Temperature changes may also take place in a region that is thin in comparison to the overall dimensions

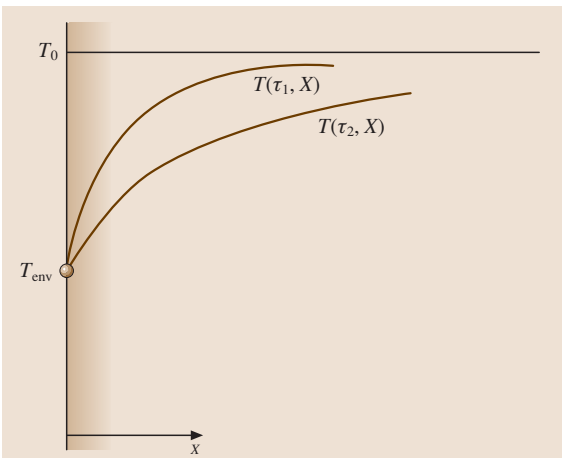


Fig. 4.34 Semi-infinite body

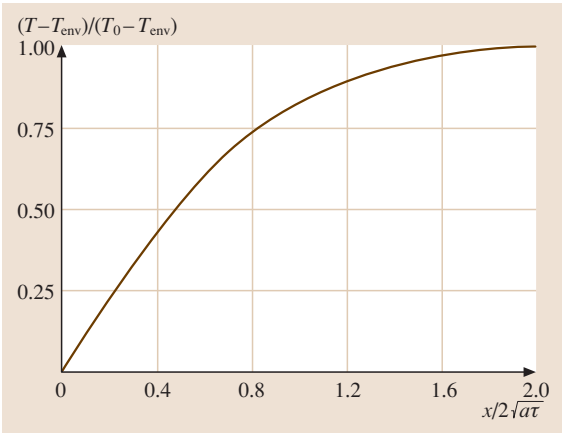


Fig. 4.35 Temperature course in a semi-infinite body

of the body. Such a body is called semi-infinite. In this case, a semi-infinite plane wall (Fig. 4.34) with a constant initial temperature T_0 is considered. At time $\tau = 0$, the surface temperature of the wall is reduced to $T(x = 0) = T_{env}$ and then remains constant. The temperature profiles at different times $\tau_1, \tau_2 \ldots$ are given by

$$\frac{T - T_{env}}{T_0 - T_{env}} = f\left(\frac{x}{2\sqrt{a\tau}}\right) \qquad (4.178)$$

with the Gaussian error function $f(x/(2\sqrt{a\tau}))$; see Fig. 4.35. The heat flux at the surface results from the differentiation $\dot{q} = -\lambda(\partial T/\partial x)_{x=0}$, which yields

$$\dot{q} = \frac{b}{\sqrt{\pi\,\tau}}(T_{env} - T_0)\,. \qquad (4.179)$$

The heat penetration coefficient $b = \sqrt{\lambda\varrho\,c}$ (SI unit $\mathrm{Ws^{1/2}/(m^2K)}$) (Table 4.25), is a measure for the heat transfer that has penetrated into the body at a given time, if the surface temperature was suddenly increased by the amount $T_{env} - T_0$ as compared to the initial temperature T_0.

Example 4.17: A sudden change in weather causes the temperature at the Earth's surface to drop from +5 °C to −5 °C. How much does the temperature decrease at a depth of 1 m after 20 days? The thermal diffusivity of the soil is $a = 6.94\times10^{-7}\ \mathrm{m^2/s}$. According to (4.178), the decrease is

$$\frac{T-(-5)}{5-(-5)} = f\left(\frac{1}{2\left(6.94\times10^{-7}\times20\times24\times3600\right)^{1/2}}\right)$$
$$= f(0.456)\,.$$

Figure 4.35 gives $f(0.456) = 0.48$, thus, $T = -0.2\,^\circ\mathrm{C}$.

Table 4.25 Heat penetration coefficients $b = \sqrt{\lambda\,\varrho\,c}$

	b ($\mathrm{Ws^{1/2}/m^2K}$)		b ($\mathrm{Ws^{1/2}/m^2K}$)
Copper	36 000	Sand	1200
Iron	15 000	Wood	400
Concrete	1600	Foam	40
Water	1400	Gases	6

Finite Heat Transfer at the Surface. According to Fig. 4.34, heat transfer is by convection from the surface of a body to the environment. At the surface, the relation $\dot{q} = -\lambda(\partial T/\partial x) = \alpha(T_w - T_{env})$ holds, with the ambient temperature T_{env} and the time-variable wall temperature $T_w = T(x = 0)$. In this case, (4.178) no longer holds. Instead, the heat transfer rate is given by

$$\dot{q} = \frac{b}{\sqrt{\pi\tau}}(T_{env} - T_0)\Phi(z)\,, \qquad (4.180)$$

where $\Phi(z) = 1 - \frac{1}{2z^2} + \frac{1\times3}{2^2z^4} - \ldots + (-1)^{n-1}\frac{1\times3\ldots(2n-3)}{2^{n-1}z^{2n-2}}$ and $z = \alpha\sqrt{a\tau}/\lambda$.

Two Semi-infinite Bodies in Thermal Contact

Two semi-infinite bodies of different, but initially constant, temperatures T_1 and T_2 with the thermal properties λ_1, a_1 and λ_2, a_2 are suddenly brought into contact at time $t = 0$ (Fig. 4.36). After a very short time at both sides of the contact area, a temperature T_m is present and remains constant. This temperature is given

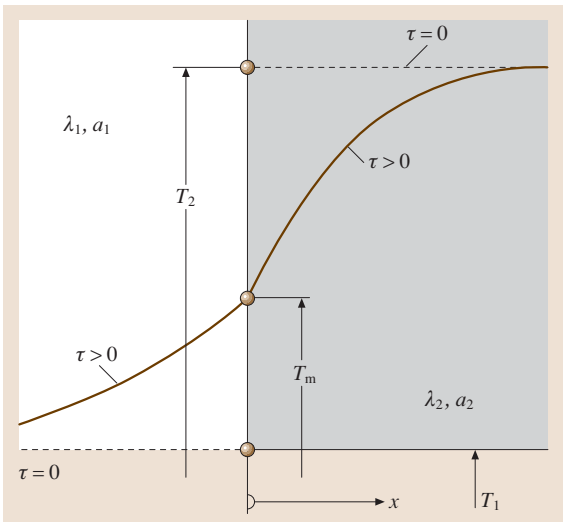


Fig. 4.36 Contact temperature T_m between two semi-infinite bodies

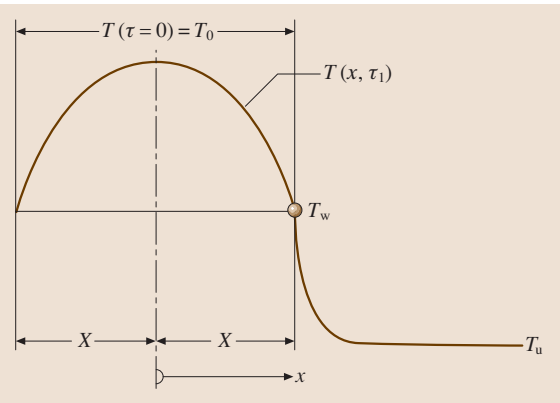


Fig. 4.37 Cooling of a flat plate

by

$$\frac{T_m - T_1}{T_2 - T_1} = \frac{b_2}{b_1 + b_2} .$$

The contact temperature T_m is closer to the temperature of the body with the higher heat penetration coefficient b. One of the values b can be determined by measuring T_m, if the other value is known.

Temperature Equalization in Simple Bodies

A simple body such as a plate, a cylinder, or a sphere may have a uniform temperature T_0 at time $\tau = 0$. Afterwards, however, it is cooled or heated due to heat transfer between the body and a surrounding fluid of temperature T_{env} given by the boundary condition $-\lambda(\partial T/\partial n)_w = \alpha(T_w - T_{env})$, where n is the coordinate perpendicular to the body surface.

Plane Plate. The temperature profile shown in Fig. 4.37 is described by an infinite series. However, for $a\tau/X^2 \geq 0.24$ (where $a = \lambda/(\varrho c)$ is the thermal diffusivity), the following relation provides a good approximation

$$\frac{T - T_{env}}{T_0 - T_{env}} = C \exp\left(-\delta^2 \frac{a\tau}{X^2}\right) \cos\left(\delta \frac{x}{X}\right) \qquad (4.181)$$

with less than a 1% error in temperature. The constants C and δ depend, according to Table 4.26, on the Biot number $\mathrm{Bi} = \alpha X/\lambda$. When $x = X$, (4.181) leads to the surface temperature T_w at the wall. The heat transfer rate follows from $\dot{Q} = -\lambda A(\partial T/\partial x)_{x=X}$.

Cylinder. The radial coordinate r replaces coordinate x in Fig. 4.37, and the radius of the cylinder is R. Again, the temperature profile is described by an infinite series, which can be approximated for $a\tau/R^2 \geq 0.21$ by

$$\frac{T - T_{env}}{T_0 - T_{env}} = C \exp\left(-\delta^2 \frac{a\tau}{R^2}\right) I_0\left(\delta \frac{r}{R}\right) \qquad (4.182)$$

with less than 1% error. The term I_0 is a Bessel function of zeroth order. Its values are presented in tables [4.22]. The constants C and δ depend, according to Table 4.27, on the Biot number. When $r = R$, the surface temperature at the wall results from (4.182) and the heat transfer rate from $\dot{Q} = -\lambda A(\partial T/\partial r)_{r=R}$, where the first derivative of the Bessel function $I_0' = I_1$ appears. The Bessel function of first order I_1 is also given in [4.23].

Sphere. The cooling or heating of a sphere of radius R is also described by an infinite series. For $a\tau/R^2 \geq 0.18$, temperature profile can be approximated by

$$\frac{T - T_{env}}{T_0 - T_{env}} = C \exp\left(-\delta^2 \frac{a\tau}{R^2}\right) \frac{\sin\left(\delta \frac{r}{R}\right)}{\delta \frac{r}{R}} \qquad (4.183)$$

with less than 2% error. The constants C and δ depend, according to Table 4.28, on the Biot number.

4.10.4 Heat Transfer by Convection

If heat transfer in fluids with bulk fluid motion is considered, in addition to (molecular) heat conduction, energy transport by convection must be taken into account. Each volume element of the fluid possesses internal energy, which is transported by the flow and, in the case considered here, is transferred by convection to a solid body.

Dimensionless Characteristic Numbers. The basis for the description of processes of convective transport is the use of similarity mechanics. These descriptions allow for the considerable reduction of the number of influencing parameters and for the expression of the general heat transfer laws for geometrically similar bodies and different substances. The following dimen-

Table 4.26 Constants C and δ in (4.181)

Bi	∞	10	5	2	1	0.5	0.2	0.1	0.01
C	1.2732	1.2620	1.2402	1.1784	1.1191	1.0701	1.0311	1.0161	1.0017
δ	1.5708	1.4289	1.3138	1.0769	0.8603	0.6533	0.4328	0.3111	0.0998

Table 4.27 Constants C and δ in (4.182)

Bi	∞	10	5	2	1	0.5	0.2	0.1	0.01
C	1.6020	1.5678	1.5029	1.3386	1.2068	1.1141	1.0482	1.0245	1.0025
δ	2.4048	2.1795	1.9898	1.5994	1.2558	0.9408	0.6170	0.4417	0.1412

Table 4.28 Constants C and δ in (4.183)

Bi	∞	10	5	2	1	0.5	0.2	0.1	0.01
C	2.0000	1.9249	1.7870	1.4793	1.2732	1.1441	1.0592	1.0298	1.0030
δ	3.1416	2.8363	2.5704	2.0288	1.5708	1.1656	0.7593	0.5423	0.1730

sionless characteristic numbers are of importance

$$\begin{aligned}
\text{Nusselt number} \quad & \mathrm{Nu} = \alpha l/\lambda \,, \\
\text{Reynolds number} \quad & \mathrm{Re} = wl/\nu \,, \\
\text{Prandtl number} \quad & \mathrm{Pr} = \nu/a \,, \\
\text{Péclet number} \quad & \mathrm{Pe} = wl/a = \mathrm{RePr} \,, \\
\text{Grashof number} \quad & \mathrm{Gr} = l^3 g\beta\Delta T/\nu^2 \,, \\
\text{Stanton number} \quad & \mathrm{St} = \alpha/\left(\varrho\, w c_\mathrm{p}\right) \\
& \quad = \mathrm{Nu}/(\mathrm{RePr}) \,, \\
\text{Geometric characteristic numbers} \quad & l_n/l;\, n = 1, 2, \ldots \,.
\end{aligned}$$

The variables signify the following: λ – thermal conductivity of the fluid, l – a characteristic length of the flow domain l_1, l_2, ..., ν – the kinematic viscosity of the fluid, ϱ – density, $a = \lambda/(\varrho\, c_\mathrm{p})$ – thermal diffusivity, c_p – constant-pressure specific heat of the fluid, g – gravitational acceleration, $\Delta T = T_\mathrm{w} - T_\mathrm{f}$ – difference between the wall temperature T_w of a cooled or heated body and the mean temperature T_f of the fluid along the body, β – thermal volume expansivity at the wall temperature with $\beta = 1/T_\mathrm{w}$ for ideal gases. The Prandtl number is a fluid property (Table 4.24).

Forced and natural convection are distinguished as follows. In forced convection, the fluid motion is caused by outer forces, e.g., by the pressure increase in a pump. In natural convection, the fluid motion is caused by density differences in the fluid and the corresponding buoyancy effects in a gravitational field. These density differences usually arise due to temperature differences, rarely due to pressure differences. In mixtures, density differences are also caused by concentration differences. The heat transfer in forced convection is described by equations of the form

$$\mathrm{Nu} = f_1(\mathrm{Re}, \mathrm{Pr}, l_n/l) \tag{4.184}$$

and in natural convection by

$$\mathrm{Nu} = f_2(\mathrm{Gr}, \mathrm{Pr}, l_n/l) \,. \tag{4.185}$$

The desired heat transfer coefficient is obtained from the Nusselt number by $\alpha = \mathrm{Nu}\lambda/l$. The functions f_1 and f_2 can be determined theoretically only for special cases. In general, they must be determined through experimentation and depend on the shape of the cooling or heating areas (even, vaulted, smooth, rough or finned), the flow structure and, usually to a minor extent, on the direction of the heat transfer (heating or cooling).

Heat Transfer Without Change of Phase

Forced Convection.

Laminar Flow Along a Flat Plate. According to *Pohlhausen* [4.24], for the mean Nusselt number of a plate of length l, the following relation holds

$$\mathrm{Nu} = 0.664\,\mathrm{Re}^{1/2}\mathrm{Pr}^{1/3} \,, \tag{4.186}$$

where $\mathrm{Nu} = \alpha l/\lambda$, $\mathrm{Re} = wl/\nu < 10^5$, and $0.6 \leq \mathrm{Pr} \leq 2000$. The material properties must be evaluated at the mean fluid temperature $T_\mathrm{m} = (T_\mathrm{w} - T_\infty)/2$, where T_w is the wall temperature and T_∞ the free-stream temperature far beyond the wall surface.

Turbulent Flow Along a Flat Plate. From about $\mathrm{Re} = 5 \times 10^5$ the boundary layer becomes turbulent. The mean Nusselt number of a plate of length l in this case is

$$\mathrm{Nu} = \frac{0.037\,\mathrm{Re}^{0.8}\mathrm{Pr}}{1 + 2.443\,\mathrm{Re}^{-0.1}(\mathrm{Pr}^{2/3} - 1)} \,, \tag{4.187}$$

where $\mathrm{Nu} = \alpha l/\lambda$, $\mathrm{Re} = wl/\nu$, $5 \times 10^5 < \mathrm{Re} < 10^7$, and $0.6 \leq \mathrm{Pr} \leq 2000$. The material properties must be evaluated at the mean fluid temperature $T_\mathrm{m} = (T_\mathrm{w} - T_\infty)/2$. T_w is the wall temperature and T_∞ the free-stream temperature far beyond the wall surface.

Flow Through Pipes in General. Below a Reynolds number of $\mathrm{Re} = 2300$ ($\mathrm{Re} = wd/\nu$, where w is the mean cross-sectional velocity and d is the pipe diameter), the flow is laminar, while above $\mathrm{Re} = 10^4$, the flow is turbulent. In the range $2300 < \mathrm{Re} < 10^4$, whether the flow

is laminar or turbulent depends on the roughness of the pipe, the means of inflow, and the shape of the pipe in the inflow section. The mean heat transfer coefficient α over the pipe length l is defined by $\dot{q} = \alpha \Delta\vartheta$, with the mean logarithmic temperature difference described by

$$\Delta\vartheta = \frac{(T_w - T_{in}) - (T_w - T_{out})}{\ln \frac{T_w - T_{in}}{T_w - T_{out}}} , \qquad (4.188)$$

where T_w is the wall temperature, T_{in} is the temperature at the inlet, and T_{out} is the temperature at the outlet cross-section.

Laminar Flow Through Pipes. A flow is termed hydrodynamically developed if the velocity profile no longer changes in the flow direction. In a laminar flow of a highly viscous fluid, the velocity profile adopts the shape of a Poiseuillean parabola after only a short distance from the inlet. The mean Nusselt number at constant wall temperature can be calculated exactly via an infinite series (the Graetz solution), which, however, converges poorly. According to *Stephan* [4.25], as an approximate solution for the hydrodynamically developed laminar flow, the following equation holds

$$Nu_0 = \frac{3.657}{\tanh(2.264X^{1/3} + 1.7X^{2/3})} + \frac{0.0499}{X}\tanh X , \qquad (4.189)$$

where $Nu_0 = \alpha d/\lambda$, $X = l/(d\,Re\,Pr)$, $Re = wd/\nu$, and $Pr = \nu/a$. This equation is valid for laminar flow ($Re \leq 2300$) in the entire range $0 \leq X \leq \infty$ and the maximum deviation from the exact values of the Nusselt number is 1%. The fluid properties must be evaluated at the mean fluid temperature $T_m = (T_w + T_B)/2$, where $T_B = (T_{in} + T_{out})/2$.

If a fluid enters a pipe at an approximately constant velocity, the velocity profile changes along the flow path until it reaches the Poiseuillean parabola after a distance described by the equation $l/(d\,Re) = 5.75 \times 10^{-2}$. According to *Stephan* [4.25], for this case, that of a hydrodynamically developed laminar flow, the following equation holds for the range $0.1 \leq Pr \leq \infty$

$$\frac{Nu}{Nu_0} = \frac{1}{\tanh(2.43\,Pr^{1/6}X^{1/6})} , \qquad (4.190)$$

where $Nu = \alpha d/\lambda$ and the quantities are defined as above. The error is less than 5% for $1 \leq Pr \leq \infty$ but is up to 10% for $0.1 \leq Pr < 1$. The fluid properties must be evaluated at the mean fluid temperature $T_m = (T_w + T_B)/2$, where $T_B = (T_{in} + T_{out})/2$.

Heat Transfer for Turbulent Flow Through Pipes. For a hydrodynamically developed flow ($l/d \geq 60$) the McAdam equation holds in the range $10^4 \leq Re \leq 10^5$ and $0.5 < Pr < 100$

$$Nu = 0.024\,Re^{0.8}Pr^{1/3} . \qquad (4.191)$$

The fluid properties have to be evaluated at the mean temperature $T_m = (T_w + T_B)/2$ with $T_B = (T_{in} + T_{out})/2$.

For hydrodynamically undeveloped flow and for developed flow, Petukhov's equation (modified by Gnielinski) holds in the range $10^4 \leq Re \leq 10^6$ and $0.6 \leq Pr \leq 1000$

$$Nu = \frac{Re\,Pr\zeta/8}{1 + 12.7\sqrt{\zeta/8}(Pr^{2/3} - 1)}\left[1 + \left(\frac{d}{l}\right)^{2/3}\right] , \qquad (4.192)$$

where the friction factor $\zeta = (0.78 \ln Re - 1.5)^{-2}$, $Nu = \alpha d/\lambda$, and $Re = wd/\nu$. The fluid properties must be evaluated at the mean temperature $T_m = (T_w + T_B)/2$. Under otherwise similar conditions, the heat transfer coefficients are larger in pipe bends than in straight pipes with the same cross section. For a pipe bend with a bend diameter D, the following equation holds, according to Hausen, for turbulent flow

$$\alpha = \alpha_{straight}\left[1 + \left(21\,Re^{0.14}\right)(d/D)\right] . \qquad (4.193)$$

A Single Pipe Placed Transversely in a Flow. The heat transfer coefficient for a pipe placed transversely in a flow can be determined from Gnielinski's equation

$$Nu = 0.3 + \left(Nu_\ell^2 + Nu_t^2\right)^{1/2} , \qquad (4.194)$$

where the Nusselt number Nu_ℓ of the laminar plate flow is described according to (4.186), Nu_t of the turbulent plate flow is described according to (4.187), and $Nu = \alpha l/\lambda$, $1 < Re = wl/\nu < 10^7$, and $0.6 < Pr < 1000$. For length l, the overflowed length $l = d\pi/2$ must be inserted. The fluid properties must be evaluated at the mean temperature $T_m = (T_{in} + T_{out})/2$. This equation holds for mean turbulence intensities of 6–10%, which can be expected in technical applications.

A Row of Pipes Placed Transversely in a Flow. Mean heat transfer coefficients for a single row of pipes placed transversely in a flow (Fig. 4.38) can also be determined using (4.194). Now, however, the Reynolds number must be calculated with the mean velocity w_m in the pipe row placed transversely in the flow.

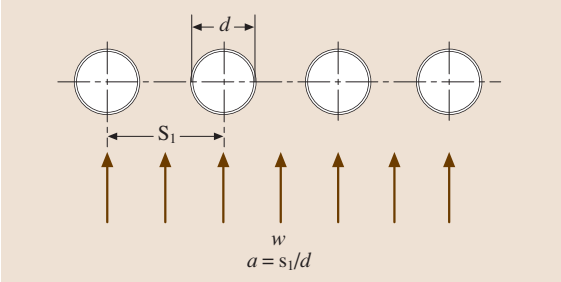


Fig. 4.38 A row of pipes placed transversely in a flow

The Reynolds number is described by the equation: $\mathrm{Re} = w_\mathrm{m} l/\nu$, where $w_\mathrm{m} = w/\psi$, w is the far-field velocity and $\psi = 1 - \pi/(4a)$ is the void space fraction, where $a = s_1/d$ (Fig. 4.38).

A Pipe Bundle. If the pipes are placed in straight lines (Fig. 4.39a), the axes of all pipes are consecutively in the flow direction. If the arrangement is staggered (Fig. 4.39b), the axes of a pipe row are shifted in comparison to the axes of the row in front. The heat transfer depends additionally on the crosswise and longwise division of the pipes, $a = s_1/d$ and $b = s_2/d$. The determination of the heat transfer coefficient starts with the calculation of the Nusselt number for a single pipe placed transversely in the flow, according to (4.194), in which the Reynolds number contains the mean velocity w_m in the pipe bundle: $\mathrm{Re} = w_\mathrm{m} l/\nu$, where $w_\mathrm{m} = w/\psi$, w is the far-field velocity of the pipe row, and ψ is the void space fraction $\psi = 1 - \pi/(4a)$ for $b > 1$ and $\psi = 1 - \pi/(4ab)$ for $b < 1$. The Nusselt number determined in this way must be multiplied by an arrangement factor f_A. This leads to the Nusselt number $\mathrm{Nu_B} = \alpha_\mathrm{B} l/\lambda$ (where $l = d\pi/2$) of the bundle

$$\mathrm{Nu_B} = f_\mathrm{A}\mathrm{Nu}\,. \tag{4.195}$$

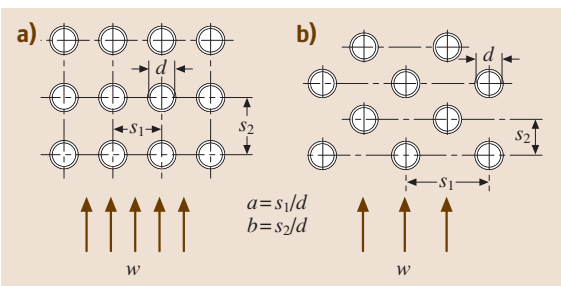


Fig. 4.39a,b Arrangement of pipes in pipe bundles: **(a)** in straight lines and **(b)** staggered

For a straight arrangement, the following holds

$$f_\mathrm{A} = 1 + 0.7(b/a - 0.3)/(\psi^{3/2}(b/a + 0.7)^2) \tag{4.196}$$

and for a staggered arrangement

$$f_\mathrm{A} = 1 + 2/(3b)\,. \tag{4.197}$$

The heat flux is $\dot{q} = \alpha\Delta\vartheta$ with $\Delta\vartheta$ according to (4.188). Equations (4.196) and (4.197) hold for pipe bundles consisting of ten or more pipe rows. For heat exchangers with fewer pipe rows, the heat transfer coefficient (4.195) must be multiplied by a factor $(1 + (n - 1) f_\mathrm{A}/n)$, where n is the number of pipe rows.

Natural Convection. The heat transfer coefficient for a vertical wall can be calculated with the equation of Churchill and Chu

$$\mathrm{Nu} = \left(\frac{0.825 + 0.387\,\mathrm{Ra}^{1/6}}{\left[1 + (0.492/\mathrm{Pr})^{9/16}\right]^{8/27}} \right)^2 , \tag{4.198}$$

in which the mean Nusselt number $\mathrm{Nu} = \alpha l/\lambda$ is formed with the wall height l, and the Rayleigh number is defined as $\mathrm{Ra} = \mathrm{GrPr}$, where the Grashof number is described by the following

$$\mathrm{Gr} = \frac{g l^3}{\nu^2} \frac{\varrho_\infty - \varrho_\mathrm{w}}{\varrho_\mathrm{w}}$$

and the Prandtl number by $\mathrm{Pr} = \nu/a$.

If natural convection is caused solely by temperature differences, the Grashof number can be written

$$\mathrm{Gr} = \frac{g l^3}{\nu^2} \beta (T_\mathrm{w} - T_\infty)\,,$$

where the volume expansivity is denoted by β, where $\beta = 1/T_\mathrm{w}$ for ideal gases. Equation (4.198) holds in the range $0 < \mathrm{Pr} < \infty$ and $0 < \mathrm{Ra} < 10^{12}$. The fluid properties must be evaluated at the mean temperature $T_\mathrm{m} = (T_\mathrm{w} - T_\infty)$. A similar equation holds according to Churchill and Chu also for natural convection in a horizontal cylinder

$$\mathrm{Nu} = \left(\frac{0.60 + 0.387\,\mathrm{Ra}^{1/6}}{\left[1 + (0.559/\mathrm{Pr})^{9/16}\right]^{8/27}} \right)^2 . \tag{4.199}$$

The same definitions used in (4.198) hold over the range of validity $0 < \mathrm{Pr} < \infty$ and $10^{-5} \le \mathrm{Ra} \le 10^{12}$, and the characteristic length is the diameter d. For *horizontal rectangular plates*, the following holds for $0 < \mathrm{Pr} < \infty$

$$\mathrm{Nu} = 0.766(\mathrm{Ra}\, f_2)^{1/5} \quad \text{if} \quad \mathrm{Ra}\, f_2 < 7 \times 10^4 \tag{4.200}$$

and

$$\mathrm{Nu} = 0.15(\mathrm{Ra}\, f_2)^{1/3} \quad \text{if} \quad \mathrm{Ra}\, f_2 > 7\times 10^4 \,, \tag{4.201}$$

where

$$f_2 = \left[1 + (0.322/\mathrm{Pr})^{11/20}\right]^{-20/11} \,,$$

where $\mathrm{Nu} = \alpha l/\lambda$, if l is the shorter side of the rectangle.

Heat Transfer in Condensation and in Boiling

Condensation. If the temperature of a wall surface is lower than the saturation temperature of adjacent vapor, the vapor is condensed at the wall surface. Depending on the wetting characteristics, the condensate forms drops or a continuous liquid film. The heat transfer coefficients are usually larger for dropwise condensation than for film condensation. However, in order to maintain dropwise condensation for a certain amount of time, particular measures such as the application of dewetting agents are necessary. Dropwise condensation therefore appears rather seldom.

Film Condensation. If the condensate flows as a laminar film on a vertical wall of height l, the mean heat transfer coefficient α is

$$\alpha = 0.943 \left(\frac{\varrho g r \lambda^3}{\nu (T_\mathrm{s} - T_\mathrm{w})} \frac{1}{l} \right)^{1/4} \,. \tag{4.202}$$

For condensation on horizontal single pipes with an outer diameter d, the following relation holds

$$\alpha = 0.728 \left(\frac{\varrho g r \lambda^3}{\nu (T_\mathrm{s} - T_\mathrm{w})} \frac{1}{l} \right)^{1/4} \,. \tag{4.203}$$

The equations require that no noticeable shear stress is exerted by the vapor on the condensate film. At Reynolds number $\mathrm{Re}_\delta = w_\mathrm{m}\delta/\nu$ (where w_m is the velocity of the condensate, d the film thickness, and ν the kinematic viscosity) between 75 and 1200 the transition to turbulent flow in the condensate film gradually takes place. In the transition range, the following relation holds

$$\alpha = 0.22\lambda/(\nu^2/g)^{1/3} \,, \tag{4.204}$$

whereas for turbulent film flow ($\mathrm{Re}_\delta > 1200$), the following relation according to Grigull holds

$$\alpha = 0.003 \left(\frac{\lambda^3 g\, (T_\mathrm{s} - T_\mathrm{w})}{\varrho \nu^3 r} l \right)^{1/2} \,. \tag{4.205}$$

Equations (4.204) and (4.205) are valid also for vertical pipes and plates but not for horizontal pipes.

Evaporation. If a liquid in a container is heated, evaporation starts after the boiling temperature T_s is exceeded. For small excess wall temperatures $T_\mathrm{w} - T_\mathrm{s}$ the liquid evaporates only on its free surface (silent boiling). Heat is transported by the buoyancy flow from the heating surface to the free surface of the liquid. For higher excess wall temperatures vapor bubbles are formed at the heating surface (nucleate boiling) and rise. They increase the movement of the liquid and thus the heat transfer. With increasing excess wall temperature, the bubbles merge more and more into a continuous vapor film, whereby the heat transfer is decreased (transition boiling). Figure 4.40 shows the different heat transfer ranges. The heat transfer coefficient α is defined as

$$\alpha = \dot{q}/(T_\mathrm{w} - T_\mathrm{s}) \,,$$

where the heat flux is $\dot{q}$ in $\mathrm{W/m^2}$.

Industrial evaporators work in the range of silent boiling or, more often, in the nucleate boiling range. In the silent boiling range the laws for heat transfer in natural convection hold, (i. e., (4.198) and (4.199)). In the nucleate boiling region

$$\alpha = c\dot{q}^n F\,(p) \;\text{ with } 0.5 < n < 0.8 \,.$$

For water at boiling pressures between 0.5 and 20 bar, according to *Fritz* [4.26], the following relation holds

$$\alpha = 1.95\dot{q}^{0.72} p^{0.24} \tag{4.206}$$

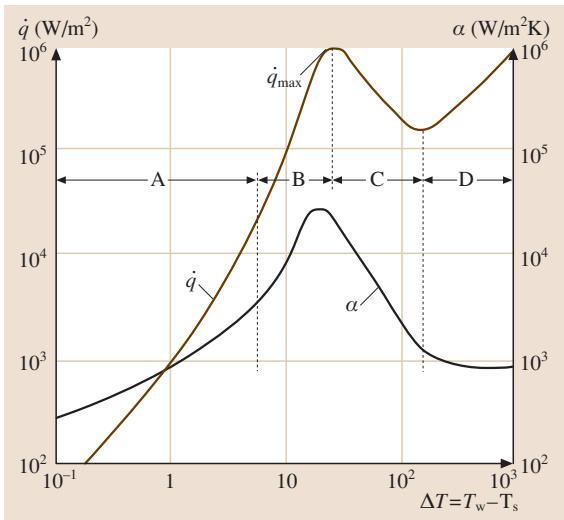


Fig. 4.40 Boiling ranges for water of 1 bar. A: natural convection (silent boiling), B nucleate boiling, C transition boiling, D film boiling

with α in W/(m^{2}K), $\dot{q}$ in W/m^2 and p in bar. According to Stephan and Preußer, for arbitrary liquids the following relation is valid for nucleate boiling close to ambient pressure

$$\mathrm{Nu} = 0.0871 \left(\frac{\dot{q}d}{\lambda' T_\mathrm{s}}\right)^{0.674} \left(\frac{\varrho''}{\varrho'}\right)^{0.156} \left(\frac{rd^2}{a'^2}\right)^{0.371} \times \left(\frac{a'^2\varrho'}{\sigma d}\right)^{0.350} (\mathrm{Pr}')^{-0.162} \ . \qquad (4.207)$$

$\mathrm{Nu} = \alpha d/\lambda'$ is formed with the detachment diameter of the vapor bubbles $d = 0.851\beta_0 \left[2\sigma/g(\varrho' - \varrho'')\right]^{1/2}$, where the contact angle is $\beta_0 = 45°$ for water, 1° for low-boiling and 35° for other liquids. Quantities denoted with a single prime relate to the boiling liquid, those with a double prime relate to the saturated vapor. The equations above are not valid for boiling in forced flow.

4.10.5 Radiative Heat Transfer

In addition to direct contact modes, heat can also be transferred by radiation. Thermal radiation (heat radiation) consists of a spectrum of electromagnetic waves in the wavelength range between 0.1 and 1000 µm. Visible light, as a reference, has a wavelength range between 0.4 and 0.76 µm. If a body is supplied with a heat transfer $\dot{Q}$ by radiation, the fraction $r\dot{Q}$ is reflected, the fraction $a\dot{Q}$ is absorbed, and the fraction $d\dot{Q}$ passes through (where $r + d + a = 1$). A body that reflects radiation completely ($r = 1,\ d = a = 0$) is called an ideal mirror, while a body that absorbs radiation completely ($a = 1,\ r = d = 0$) is called a black body. A body is called diathermal ($d = 1,\ r = a = 0$) if radiation passes completely through, where examples for this are gases such as O_2, N_2, etc.

Stefan–Boltzmann Law

Every body emits radiation corresponding to its surface temperature. The maximum radiation possible is emitted by a black body. It can be experimentally approximated by a blackened surface (e.g., with soot) or by a hollow space, whose walls have the same temperature everywhere, that has a small opening to let radiation out. The total radiation emitted by a black body per unit area is

$$\dot{e}_\mathrm{s} = \sigma T^4 \ , \qquad (4.208)$$

where $\dot{e}_\mathrm{s}$ is called the emission (W/m^2) of the black radiator, and $\sigma = 5.67 \times 10^{-8}$ W/m^{2}K^4 is the *radiation coefficient*, also called the Stefan–Boltzmann constant. The emission $\dot{e}_\mathrm{s}$ is an energy flux and thus equal to the heat flux $\dot{q}_\mathrm{s} = \mathrm{d}\dot{Q}/\mathrm{d}A$ a black radiator emits. With the emission $\dot{e}_\mathrm{n}$ in a normal direction and $\dot{e}_\varphi$ in the direction of angle φ to the normal, *Lambert's cosine law* $\dot{e}_\varphi = \dot{e}_\mathrm{n} \cos\varphi$ for black radiators holds true. Often the radiation of real bodies differs from this general law, however.

Kirchhoff's Law

Real bodies emit less than black radiators, where the energy emitted from real surfaces is

$$\dot{e} = \varepsilon \dot{e}_\mathrm{s} = \varepsilon\sigma T^4 \qquad (4.209)$$

with the emissivity being in the range $0 < \varepsilon < 1$ and in general depending on temperature (Table 4.29). In limited temperature ranges, many engineering surfaces (with the exception of shiny metal) can be interpreted as grey radiators. The energy radiated by them is distributed over wavelength in the same way as it is for black radiators, but reduced by a factor $\varepsilon < 1$. Strictly speaking, $\varepsilon = \varepsilon(T)$ holds true for grey radiators. For small temperature ranges, however, it is admissible to assume ε as constant. Assuming a body emitts the energy per unit area $\dot{e}$, and this energy flux strikes another body, this second body absorbs the energy or rather the heat transfer

$$\mathrm{d}\dot{Q} = a\dot{e}\ \mathrm{d}A \ . \qquad (4.210)$$

The absorptivity defined by this equation depends on the temperature T of the origin of the incident radiation and on the temperature T' of the receiving surface. For black bodies, this value is $a = 1$, as all radiation striking the surface is absorbed. For surfaces which are not black, this value is $a < 1$. For grey radiators, the absorptivity is $a = \varepsilon$. According to Kirchhoff's law, the emissivity is equal to the absorptivity, $\varepsilon = a$, for each surface which is in thermal equilibrium with its environment so that the temperature of the surface does not change in time.

Heat Exchange by Radiation

Between two parallel black surfaces of temperatures T_1 and T_2 and area A, which is very large in comparison to their separation, the heat transfer

$$\dot{Q}_{12} = \sigma A\left(T_1^4 - T_2^4\right) \qquad (4.211)$$

is exchanged by radiation. Between grey radiators with emissivities ε_1 and ε_2, the heat transfer is

$$\dot{Q}_{12} = C_{12} A\left(T_1^4 - T_2^4\right) \qquad (4.212)$$

Table 4.29 Emissivity ε at temperature t

Substance	Surface	t °C	ε
Roofing paper		21	0.91
Oak wood	Planed	21	0.89
Enamel varnish	Snow white	24	0.91
Glass	Smooth	22	0.94
Lime mortar	Rough, white	21–83	0.93
Marble	Light grey, polished	22	0.93
Porcelain	Glazed	22	0.92
Soot	Smooth	–	0.93
Chamotte slab	Glazed	1000	0.75
Spirit varnish	Black, shiny	25	0.82
Brick	Red, rough	22	0.93–0.95
Water	Vertical radiation	–	0.96
Oil	Thick layer	–	0.82
Oil coating		–	0.78
Aluminum	Rough	26	0.071–0.087
Aluminum	Polished	230	0.038
Lead	Polished	130	0.057
Gray cast iron	Turned off	22	0.44
Gray cast iron	Liquid	1330	0.28
Gold	Polished	630	0.035
Copper	Polished	23	0.049
Copper	Rolled	–	0.16
Brass	Polished	19	0.05
Brass	Polished	300	0.031
Brass	Dead	56–338	0.22
Nickel	Polished	230	0.071
Nickel	Polished	380	0.087
Silver	Polished	230	0.021
Steel	Polished	–	0.29
Zinc	Zinc-coated iron sheet	28	0.23
Zinc	Polished	230	0.045
Zinc	Shiny, tinned sheet	24	0.057–0.087
Oxidized metals			
Iron	Red, slightly rosted	20	0.61
Iron	Totally rusted	20	0.69
Iron	Smooth or rough cast skin	23	0.81
Copper	Black	25	0.78
Copper	Oxidized	600	0.56–0.7
Nickel	Oxidized	330	0.40
Nickel	Oxidized	1330	0.74
Steel	Dead oxidized	26–356	0.96

with the *radiation exchange number*

$$C_{12} = \sigma / \left(\frac{1}{\varepsilon_1} + \frac{1}{\varepsilon_2} - 1 \right) . \tag{4.213}$$

Between an internal pipe with outer surface A_1 and an external pipe with inner surface A_2, which are both grey radiators with emissivities ε_1 and ε_2, respectively, the heat transfer rate is given according to (4.212), however, with

$$C_{12} = \sigma / \left[\frac{1}{\varepsilon_1} + \frac{A_1}{A_2} \left(\frac{1}{\varepsilon_2} - 1 \right) \right] . \tag{4.214}$$

If $A_1 \ll A_2$, e.g., for a pipe in a large room, it holds that $C_{12} = \sigma\varepsilon_1$. Between two surfaces of areas A_1, A_2, temperatures T_1, T_2, and emissivities ε_1, ε_2, which are arbitrarily arranged in space, a heat flow

$$\dot{Q}_{12} = \frac{\varepsilon_1 \varepsilon_2 \varphi_{12}}{1 - (1 - \varepsilon_1)(1 - \varepsilon_2)\varphi_{12}\varphi_{21}} \sigma A_1 \left(T_1^4 - T_2^4 \right) \tag{4.215}$$

exists, where φ_{12} and φ_{21} are the so-called view factors that depend on the geometric arrangement or the surfaces, values of which are given in [4.27].

Gas Radiation

Most gases are transparent to thermal radiation and neither emit nor absorb radiation. Exceptions are carbon dioxide, carbon monoxide, hydrocarbons, water vapor, sulfur dioxide, ammonia, hydrochloric acid, and alcohols. They emit and absorb radiation only in certain wavelength regions. The emissivity and absorptivity of these gases depend not only on temperature, but also on the geometric shape of the gas body.

References

4.1 F. Pavese, G.F. Molinar: *Modern Gas-Based Temperature and Pressure Measurements* (Plenum, New York 1992)

4.2 O. Knoblauch, K. Hencky: *Anleitung zu genauen technischen Temperaturmessungen*, 2nd edn. (Oldenbourg, München 1926)

4.3 VDI/VDE (Ed.): *Temperature Measurement in Industry - Principles and Special Methods of Temperature Measurement*, VDI/VDE 3511 (VDI/VDE-Gesellschaft Mess- und Automatisierungstechnik, Berlin 1996)

4.4 D. Rathmann, J. Bauer, P.A. Thompson: *A Table of Miscellaneous Thermodynamic Properties for Various Sustances, with Emphasis on the Critical Properties* (Max-Planck-Inst. Strömungsforsch., Göttingen 1978), Ber. 6

4.5 N.E. Holden, R.L. Martin: Atomic weights of elements 1981, Pure Appl. Chem. **55**, 1102–1118 (1983)

4.6 D. Ambrose: *Vapour-Liquid Critical Properties* (Nat. Phys. Lab., Teddington 1980)

4.7 K. Schäfer, G. Beggerow (Eds.): *Mechanical-Thermal Properties of State*, Landolt-Börnstein, Vol. II/1, 6th edn. (Springer, Heidelberg 1971) pp. 245–297

4.8 J.R. Dymond, E.B. Smith: *The Virial Coefficients of Pure Gases and Mixtures* (Clarendon, Oxford 1980)

4.9 R.C. Reid, J.M. Prausnitz, B.E. Poling: *The Properties of Gases and Liquids*, 4th edn. (McGraw-Hill, New York 1986)

4.10 W. Wagner, A. Kruse: *Properties of Water and Steam. Zustandsgrößen von Wasser und Wasserdampf* (Springer, Heidelberg 1998)

4.11 H.D. Baehr, K. Schwier: *Die thermodynamischen Eigenschaften der Luft* (Springer, Berlin 1961), in German

4.12 R. Span, W. Wagner: Equations of state for technical applications, III. Results for polar fluids, Int. J. Thermophys. **24**, 111–162 (2003)

4.13 R.C. Wilhoit, B.J. Zwolinski: *Handbook of Vapor Pressures and Heats of Vaporization of Hydrocarbons and Related Compounds*, Thermodyn. Res. Center Dept. Chem. Texas A&M Univ. (American Petroleum Institute Research, Texas 1971), Publ. 101, Proj. 44

4.14 R. Tillner-Roth, F. Harms-Watzenberg, H.D. Baehr: Eine neue Fundamentalgleichung für Ammoniak, DKV-Tagungsbericht **20**(II/1), 167–181 (1993)

4.15 R. Span, W. Wagner: A new equation of state for carbon dioxide covering the fluid region from the triple-point temperature to 1100 K at pressures up to 800 MPa, J. Phys. Chem. Ref. Data **25**, 1509–1596 (1996)

4.16 R. Tillner-Roth: Die thermodynamischen Eigenschaften von R134a, R152a und ihren Gemischen – Messungen und Fundamentalgleichungen, Forsch.-Ber. DKV (1993)

4.17 R. Tillner-Roth, H.D. Baehr: An international standard formulation for the thermodynamic properties of 1,1,1,2-tetrafluoroethane (HFC-134a) for temperatures from 170 K to 455 K and pressures up to 70 MPa, J. Phys. Chem. Ref. Data **23**, 657–729 (1994)

4.18 W. Wanger, V. Marx, A. Pruß: A new equation of state for chlorodifluoromethane (R22) covering the entire fluid region from 116 K to 550 K at pressures up to 200 MPa, Int. J. Refrig. **16**, 373–389 (1993)

4.19 F. Brandt: *Brennstoffe und Verbrennungsrechnung*, 3rd edn. (Vulkan, Essen 1999), in German

4.20 H.D. Baehr: Zur Thermodynamik des Heizens, Part I, Brennst. Wärme Kraft **32**, 9–15 (1980), in German

4.21 E. Schmidt: *Properties of Water and Steam in SI Units*, 3rd edn. (Springer, Berlin 1982)
4.22 I.N. Bronstein: *Taschenbuch der Mathematik*, 5th edn. (Deutsch, Frankfurt/Main 2000), in German
4.23 I.N. Bronshtein, K.A. Semendyayev, G. Musiol, H. Mühlig: *Handbook of Mathematics*, 5th edn. (Springer, Berlin 2007)
4.24 E. Pohlhausen: Der Wärmeaustausch zwischen festen Körpern und Flüssigkeiten mit kleiner Reibung und kleiner Wärmeleitung, Z. Angew. Math. Mech. **1**, 115–121 (1921)
4.25 H.D. Baehr, K. Stephan: *Heat and Mass Transfer* (Springer, Berlin 2006)
4.26 W. Fritz: *In VDI-Wärmeatlas* (VDI, Düsseldorf 1963), Hb2
4.27 VDI/GVC (Ed.): *VDI-Wärmeatlas*, 10th edn. (Springer, Berlin 2006), in German

5. Tribology

Ludger Deters

The main subjects of this chapter are the tribotechnical system, friction, wear and lubrication. Regarding the tribotechnical system essential information on structure, real contact geometry, tribological loads, operating and loss variables are provided. Concerning friction the different friction types, states and mechanisms are discussed. In the sections on wear a lot of details on types and mechanisms of wear, wear profiles and the determination of wear and the average useful life are introduced. The sections on lubrication contain relevant expositions on the lubrication states, like hydrodynamic, elastohydrodynamic, hydrostatic, mixed and boundary lubrication and lubrication with solid lubricants, on the lubricants, like mineral, synthetic and biodegradable oils and additives, lubricating greases and solid lubricants, and on the properties of lubricants, like the behaviour of the oil viscosity depending on temperature, pressure and shear rate and the consistency of lubricating greases.

5.1 Tribology

Tribology is the science and technology of interacting surfaces in relative motion. Tribology includes boundary-layer interactions both between solids and between solids and liquids and/or gases. Tribology encompasses the entire field of friction and wear, including lubrication [5.1].

Tribology aims to optimize friction and wear for a particular application case. Apart from fulfilling the required function, this means assuring high efficiency and sufficient reliability at the lowest possible manufacturing, assembly, and maintenance costs.

Friction and wear are frequently undesirable. While friction impairs the efficiency of machine elements, machines, and plants and thus increases the energy demand, wear diminishes the value of components and assemblies and can lead to the failure of machines and plants. On the other hand, many technical applications strive for high friction, e.g., brakes, clutches, wheels/rails, car tires/road, friction gears, belt drives, bolted joints, and press fits. To a limited extent, wear can also be advantageous in special cases, e.g., in breaking-in processes.

Friction and wear are not properties specific to the geometry or substance of only one of the elements involved in friction and wear, e.g., external dimensions, surface roughnesses, thermal conductivity, hardness, yield point, density or structure, but rather are properties of a system. The system's friction and/or wear behavior can already change seriously when one influencing variable of the tribotechnical system is marginally modified.

Lubrication is employed to lessen friction and minimize wear or to prevent them entirely. In the case of circulatory lubrication, the lubricant can additionally remove wear particles and heat from the friction contact. Other important tasks of lubrication are preventing corrosion (rusting) and, in the case of grease lubrication, sealing the friction points.

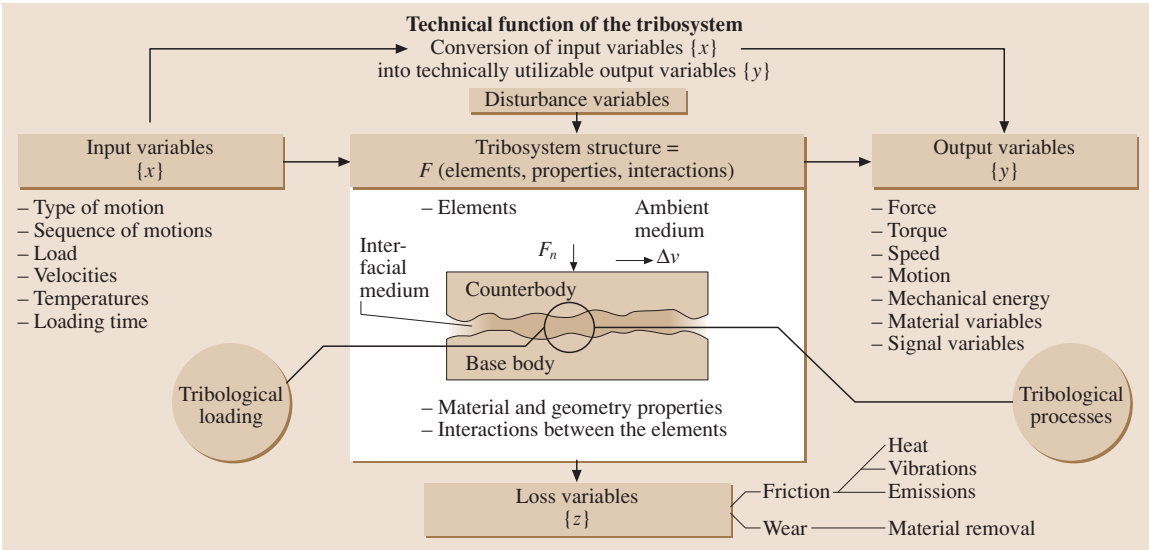


Fig. 5.1 Expanded representation of a tribotechnical system (TTS) (after [5.2])

5.1.1 Tribotechnical System

General Description

Friction and wear occur within a tribotechnical system (TTS). To delimit a TTS, a system envelope is appropriately placed around the components and materials directly involved in friction and wear, thus virtually isolating these from the remaining components. The materials and components involved in friction and wear are the elements of the TTS and are characterized by their material and shape properties. A tribotechnical system is described by the function to be fulfilled, the input variables (operating variables), the output variables, the loss variables, and the structure (Fig. 5.1).

Apart from desired *input variables*, undesired input variables, so-called *disturbance variables*, also arise. Together with the structure, they influence the *output* and *loss variables* of the TTS.

The function of a TTS is to use the system structure to convert input variables (e.g., input torque, input speed, input type of motion, and sequence of motions) into technically utilizable output variables (e.g., output torque, output speed, output motion) (Fig. 5.1).

Structure

The elements involved, their properties, and the interactions between the elements describe the structure of a TTS. The basic structure of all TTS consists of four elements: the base body, counterbody, interfacial medium, and ambient medium (Fig. 5.1). Table 5.1 displays some TTS with different elements. While the base body and counterbody are found in every TTS, the interfacial medium and, in a vacuum, even the ambient medium can be absent.

In transport and machining processes, the base body is constantly stressed by new material zones of the counterbody. Such systems are called *open* TTS. By contrast, the stressed zones of the base body and counterbody in *closed* TTS are repeatedly in contact. Examples of open and closed systems can also be found in Table 5.1. The function in open systems mainly depends on the wear of the base body. The counterbody generates the load; as a rule, the wear on it is not of interest. By comparison, when systems are closed, the ability to operate depends on the wear of both friction bodies. The elements of the TTS are characterized by a large number of *properties*, largely listed in Table 5.1.

A difference is made principally between geometry and material properties in the base body and counterbody, which are supplemented by physical variables. Interfacial medium and ambient medium can appear in different aggregate states, on which other important tribological properties depend. A difference in the material properties of the base body and counterbody is made between bulk material and the near-surface zone. The properties of the near-surface zone, e.g., structural composition, hardness, and chemical composition, are particularly important for the tribological processes. In addition, the surface roughnesses play an important role.

Table 5.1 Examples of elements of tribotechnical systems

TTS	Base body	Counterbody	Interfacial medium	Ambient medium	System type
Press and shrink joints	Shaft	Hub	–	Air	Closed
Sliding bearing	Journal	Bearing bush	Oil	Air	Closed
Mechanical face seal	Seal head	Seat	Liquid or gas	Air	Closed
Gear train	Pinion	Wheel	Gear oil	Air	Closed
Wheel/rail	Wheel	Rail	Moisture, dust, grease	Air	Open
Excavator bucket/ excavated material	Bucket	Excavated material	–	Air	Open
Turning tool	Cutting edge	Workpiece	Cutting lubricant	Air	Open

Figure 5.2 shows a diagram of the possible composition of *boundary layers* in metallic materials. The undisturbed *basic structure* generally has attached to it a hardened composition of machined or deformed layers endowed with a structure that is fine-grained compared with the basic structure. A reaction layer and an adsorption layer lie on top of this. Taken together they are also called the *outer boundary layer*. Wear is generally acceptable as long as it occurs in the outer boundary layer.

As a rule, not only the near-surface zone's material structure but also its chemical composition differs from those the bulk material. This is apparent in Fig. 5.3. The material beneath the surface already changes during manufacturing in terms of the concentration of elements present compared with the bulk material. Other considerable changes undergone by the near-surface zone's element concentrations are caused by breaking-in or occur after a short running time.

Contact Geometry

Not only the input variables, the material properties and the physical variables of the base body and counterbody, the interfacial medium, and the ambient medium but also the contact areas greatly influence friction and wear of the base body and counterbody and the tribotechnical system's lubrication state.

In turn, the contact areas appearing during operation depend on the form of contact (Table 5.2)), the input variables, the geometric properties of the base body and counterbody (Table 5.1), and the other system properties.

In the contact areas, a difference is made between the *nominal or apparent contact area* and the *real contact area* (Fig. 5.4).

The nominal or apparent contact area A_a corresponds to the macroscopic contact area of the bodies in contact, e.g., of the contact area *ab* of a cuboid on one plane or the Hertzian contact area between a cylin-

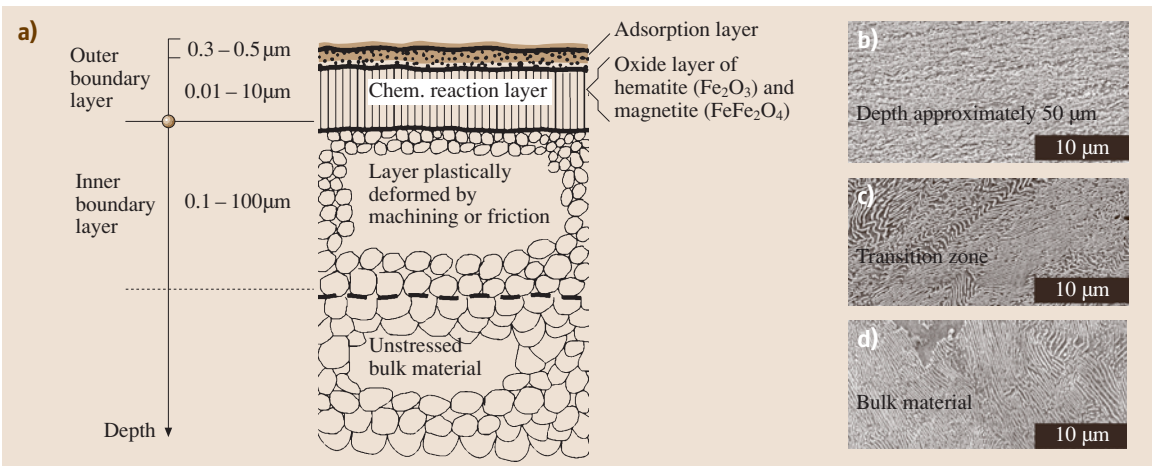


Fig. 5.2a–d Boundary-layer composition in metallic materials using a tribologically loaded rail steel (after [5.3])

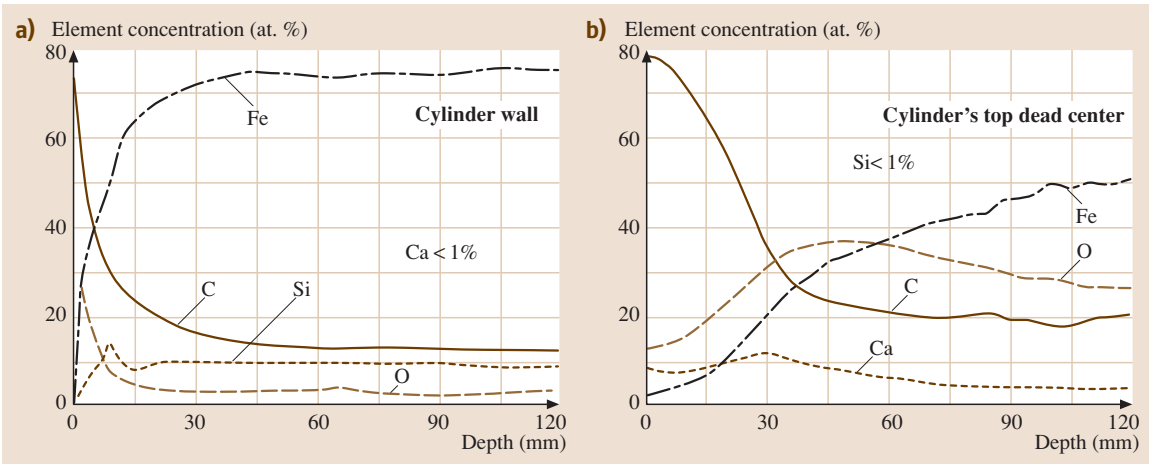


Fig. 5.3a,b Auger electron spectroscopy depth profiles as element concentration in the cylinder's top dead center of a diesel engine for new condition **(a)** before running in and **(b)** after 15 h running time (oil: SAE 15 W 40) [5.4]

der and a plane. Not only the apparent contact area but also the real contact areas $A_{r,i}$ play a key role when the asperities of the friction bodies meet in the apparent contact area.

The real contact areas result from the asperity contacts, which occur when the friction bodies are not completely separated by a lubricating film or in cases of application in which no lubricant is used (dry friction). Not only the roughness but also the waviness on the friction body surfaces have to be taken into account when analyzing the real contact areas. The waviness causes so-called *contour areas* $A_{c,i}$ to form as well as the real contact areas on the asperity contacts within the contour areas.

As a rule, the total of the real contact areas A_r, which is dependent on the roughness distributions and the separation of the two friction body surfaces, is substantially smaller than the apparent contact area ($A_r \approx 10^{-1}$–$10^{-4} A_a$). Hence the real contact pressures in the asperity contacts are substantially higher than the nominal pressure. While the nominal pressure displays elastic macromaterial behavior, plastic deformation may already have begun for a majority of the microcontacts (real contact areas). Results of calculations have yielded that the total of the real contact areas is nearly proportional to the normal force F_n [5.5]. In addition, as the normal force increases, the number of real individual contacts increases, while the real individual contact area $A_{r,i}$ remains roughly constant.

Apart from the real contact area, the *overlap ratio* ε also plays an important role as well. It represents the ratio of the apparent contact area A_a to a friction body's friction area A_f. Thus, for example, the nonrotating bush of a sliding bearing with bearing clearance has an overlap ratio of $\varepsilon = 1$, since the apparent contact area A_a for the bearing bush corresponds to the friction area A_f. For the rotating shaft however, the friction area $A_f = \pi db$, with shaft diameter d and bearing bush width b, is larger than the apparent contact area $A_a = db\gamma/2$, with contact angle γ, so that the overlap ratio is $\varepsilon < 1$. For a constantly loaded friction body, an overlap ratio $\varepsilon = 1$ means permanent contact, no cyclical mechanical loads (macroscopic), permanent frictional heat absorption, and limited microchemical reaction with the ambient medium. For the friction body concerned, an overlap ratio of $\varepsilon < 1$ leads to an intermittent contact, to cyclical mechanical loading, to intermittent frictional heat absorption, and to tribochemical reactions with the ambient medium in the range A_f–A_a. When both friction bodies (the base body and counterbody) have an overlap ratio of $\varepsilon \approx 1$, the wear particles can remain in the contact area and hence adversely influence the further wear profile.

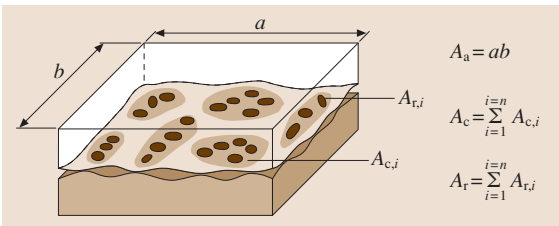


Fig. 5.4 Various types of contact areas

Table 5.2 Tribologically relevant properties of elements of the tribotechnical system (TTS) (Fig. 5.1)

1. Base body and counterpart

1.1 *Geometric poperties*

- External dimensions
- Shape and position tolerances
- Waviness
- Surface roughnesses

1.2 *Material properties*

1.2.1 Bulk material

- Strength
- Hardness (macro, micro, and Martens hardness)
- Structure, texture, microstructure phases (distribution, size, number type)
- Modulus of el. Poisson's ratio
- Residual stress
- Chemical composition

1.2.2 Near-surface zone

- Hardness (macro, micro, and Martens hardness)
- Surface energy
- Metallurgical structures, texture, microstructure phases (distribution, size, number type)
- Chemical composition
- Modulus of el. Poisson's ratio
- Residual stress
- Boundary-layer thickness and structure

1.3 Physical variables

- Density
- Heat conductivity
- Coefficient of thermal expansion
- Melting point
- Spec. thermal capacity
- Hygroscopic properties

2. Interfacial medium (lubricant)

- Aggregate state (solid, liquid, gaseous)
- For *solid* interfacial medium
  - Hardness
  - Grain size distribution
  - Grain shape
  - Grain quantity, grain number
  - Number of components, mixing ratio
  - Chemical composition
- For *liquid* interfacial medium
  - Viscosity depending on temperature, pressure, shear rate
  - Consistency
  - Wettability
  - Lubricant quantity and pressure
  - Chemical composition
  - Mixing ratio of components

3. Ambient medium

- Aggregate state (solid, liquid, gaseous)
- Heat conductivity
- Chemical composition
- Moisture
- Ambient pressure

Tribological Loads and Interactions

Tribological loads in a TTS are generated by the input and disturbance variables' action on the system structure. They chiefly include contact, kinematic, and thermal processes [5.2]. According to [5.1], the tribological load represents "the loading of the surface of

a solid caused by contact and relative motion of a solid, liquid or gaseous counterbody." It is introduced via the real contact areas. Plastic deformation and wear can cause the real contact areas to change during TTS operation.

When mechanical energy is converted by friction, energy dissipates, which makes itself noticeable by changing the thermal situation. Since the thermal behavior also continuously adapts to the new conditions as a result of wear, changes to the contact geometry, and resulting changes in the friction, dynamic rather than static influencing variables determine the tribological loading in a real contact.

The contact geometry, the processes occurring in the contact, and the thermal behavior of a TTS are influenced by, among other things, the load, the motion conditions, the element properties, and the friction state.

While the apparent contact area alone is decisive in fluid lubrication, according to [5.6], in mixed lubrication, i. e., when the dimensionless film parameter

$$\Lambda = \frac{h_{\min}}{\left(R_{q1}^2 + R_{q2}^2\right)^{1/2}}\,, \tag{5.1}$$

with the minimum lubrication film thickness $h_{\min}$ and the root-mean-square (rms) surface roughnesses R_{q1} and R_{q2} of the base body and counterbody is in the range $\Lambda < 3$, in boundary lubrication with $\Lambda < 1$ and for dry friction both the apparent contact area and the real contact areas must be allowed for (Fig. 5.4).

When there are contacts between the friction bodies, *interactions* occur in the real contact areas and in the near-surface zones. *Atomic/molecular* interactions occur on the one hand and *mechanical* interactions on the other. Whereas the former cause adhesion on solid–solid boundary layers or are extremely important technically in the form of physisorption and chemisorption on solid–fluid boundary layers, the latter lead to elastic and plastic contact deformations and to the development of the real contact areas.

The type of interaction that primarily occurs depends greatly on the friction state. Thus, when a lubricant is present the atomic/molecular interaction can be disregarded more often than the mechanical.

Friction and wear in a given TTS ultimately depend on the interactions between the elements. The friction state, the effective mechanisms of friction and wear, and the contact state can be used to describe the interactions.

The tribological loads occurring in the real contact areas produce *tribological processes*. These subsume the dynamic physical and chemical mechanisms of friction and wear and boundary-layer processes that can be attributed to friction and wear.

Operating Variables (Input Variables)

According to [5.1], the operating variables are: type of motion, the time sequence of motions of the elements contained in the system structure, and a number of technical-physical load parameters, which act on the system structure when the function is executed. The operating variables originate from:

- Type of motion and time sequence of motions
- Load
- Velocities
- Temperatures
- Loading time

The type of motion can frequently be attributed to one of the basic types of motion *sliding*, *rolling*, *spin*, *impact* or *flowing* or can be composed from these. The time sequence of motions can occur regularly, irregularly, back and forth, or intermittently. The sequence of motions frequently also consists of different components. As a rule, the normal force F_n is decisive for the load.

Both the relative velocity between the friction bodies and the entraining velocity of the lubricant in the contact and the slippage as a ratio of the relative velocity to the average circumferential velocity play a role for the velocities. The friction body temperatures and the effective contact temperature produced in operation are critically important for the temperature variables. It is normally not possible to measure the contact temperatures. Apart from these desired input variables, which as a rule are specified by a technical function, *disturbance variables* such as vibrations or dust particles must be considered under certain circumstances.

Output Variables (Useful Variables)

The TTS provides output variables for subsequent utilization. These useful variables reflect the performance of a function of the TTS. The useful variables can differ over extremely wide ranges depending on the main task of the TTS. In an energy-determined system, for example, the following output variables may be desired:

- Force
- Torque
- Velocity
- Motion
- Mechanical energy

Particular material or signal variables could be interesting as useful variables in a material- or signal-determined TTS.

Loss Variables

The loss variables of a TTS are essentially represented by friction and wear. While friction leads to losses of force, torque or energy, wear means a progressive loss of material.

The energy losses produced when there is friction are converted into heat for the most part. This process is irreversible and is called energy dissipation. Along with the conversion of friction into heat and the generation of wear particles, the tribological process generates other tribologically induced loss variables such as vibrations that frequently become apparent through sound waves, photon emission (triboluminescence), electron, ion emission, etc.

5.1.2 Friction

General

Friction can be ascribed to the interactions between bodies' material zones that are in contact or moving relative to one another; it counteracts relative motion. External and internal friction are differentiated. When friction is external, the different friction bodies' material zones are in contact; when friction is internal, material zones that are in contact belong to one friction body or the interfacial medium.

A number of parameters can characterize friction. Thus, depending on the application, friction is characterized by the friction force F_f, the friction torque M_f or the coefficient of friction f. Instead of f the symbol μ is also frequently used for the coefficient of friction. The coefficient of friction f is formed from the ratio of the friction force F_f to the normal force F_n

$$f = \frac{F_f}{F_n} . \tag{5.2}$$

The work of friction or friction energy W_f is used to calculate the frictional heat or the amount of deformation of the friction force in solid friction. It is calculated as

$$W_f = F_f s_f; , \tag{5.3}$$

with the friction distance s_f. The friction power P_f is of interest for an energy balance or efficiency calculation. The friction power is a power loss and, disregarding signs, the following applies

$$P_f = F_f \Delta v, \tag{5.4}$$

with the relative velocity Δv. (The power loss is frequently defined negatively).

Types of Friction

Friction can be classified according to various features. Types of Friction are distinguished depending on the type of relative motion between the friction bodies. Figure 5.5 presents the most important types of friction with samples applications. There are three main types of friction:

- Sliding friction
- Rolling friction
- Spin friction

Apart from these three kinematically defined types of friction, there can be overlaps (mixed forms), namely:

- Sliding–rolling friction (rolling friction)
- Sliding–spin friction
- Rolling–spin friction

Along with the types of friction shown in Fig. 5.5, another type of friction is impact friction, which applies when a body strikes another body perpendicular or oblique to the contact surface and possibly withdraws again. The angular contact ball bearing is a machine element in which sliding and rolling and spin friction appear.

Friction States

Various friction states can be defined if friction is classified as a function of the aggregate state of the material zones involved. To illustrate this, Fig. 5.6 presents different states of friction based on the Stribeck curve using a radial sliding bearing as an example. Generally, the following friction states are differentiated:

- Solid friction
- Mixed friction
- Fluid friction
- Gas friction

In *solid friction* the friction acts between material zones that exhibit solid properties and are in direct contact. If the friction occurs between solid boundary layers with modified properties compared with the bulk material, e.g., between reaction layers, then this is *boundary-layer friction*. If the boundary layers on the contact surfaces each consist of a molecular film coming from a lubricant, then this is called *boundary*

friction. In boundary friction, the lubricant's hydrodynamic effect can be disregarded because the velocity is very low and/or only a very small quantity of lubricant, insufficient to fill the lubrication gap, is present.

Fluid friction is internal friction in the lubricating film between the friction body surfaces, the surfaces being completely separated by the lubricating film. A difference is frequently made between fluid friction in conformal contact surfaces (hydrodynamics) and in nonconformal contact surfaces (elastohydrodynamics). While rigid surfaces and lubricant viscosity only dependent on temperature are generally assumed in the first case, this cannot be assumed in the second case, in which not only the deformations of the surfaces but also the lubricant viscosity's dependence on pressure, temperature, and shear rate must be taken into account.

Mixed friction is a mixed form of the friction states, boundary friction and fluid friction to be precise.

Friction Mechanisms

Solid friction can be ascribed to interactions between the elements. As already addressed before, there are essentially two different types of interaction, i. e., atomic/molecular and mechanical. *Kragelski* [5.7] speaks of friction's 'dual nature.' Hence, the friction mechanisms can be divided into two groups. Generally a difference can initially be made between the following four friction mechanisms, compiled schematically in Fig. 5.7:

- Adhesive bond shearing
- Plastic deformation
- Abrasion
- Hysteresis losses in elastic deformation

Adhesion is an atomically/molecularly based friction mechanism. Its frictional effect is based on the bonds formed atomically or molecularly in the real contact areas that separate again when relative motion occurs, as a result of which energy loss occurs.

Deformation, abrasion, and hysteresis can be classified as mechanically based friction mechanisms. The action of friction in deformation and abrasion can above all be ascribed to the displacement of overlaps of asperities. Hysteresis is based on internal friction and

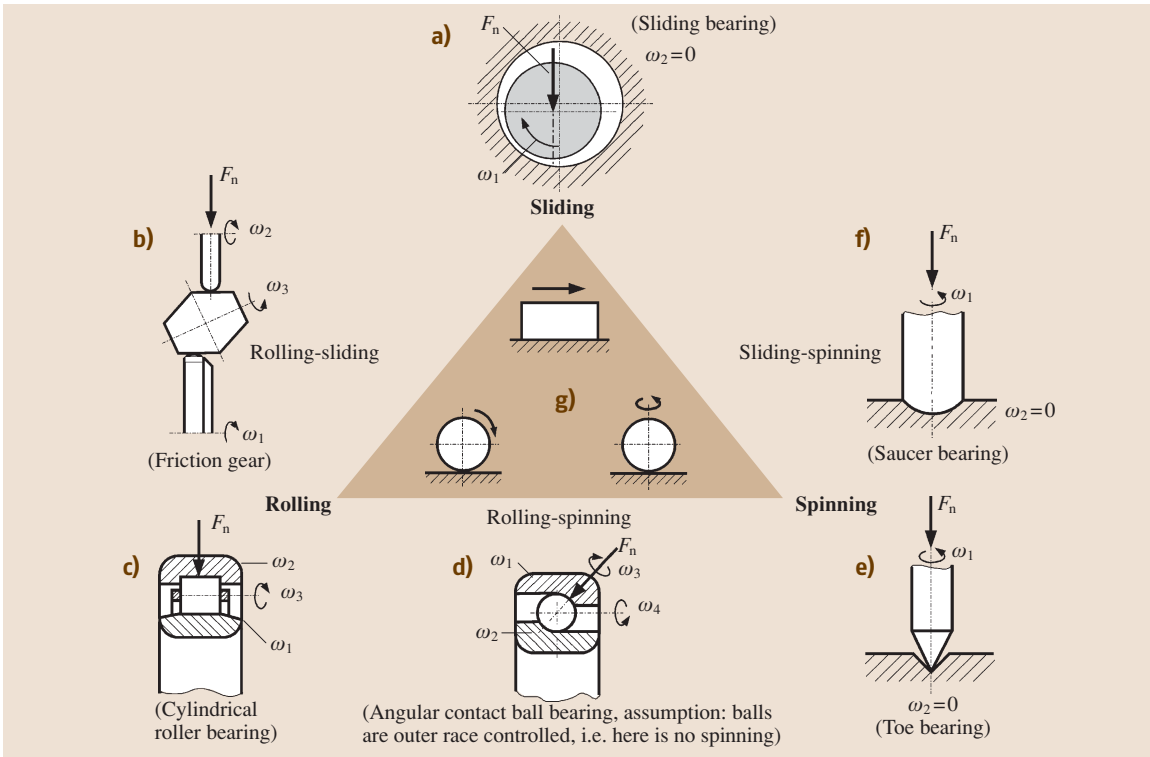


Fig. 5.5a–g Types of friction in dynamic friction

has a damping effect. Different friction mechanisms frequently appear at the same time. The friction mechanisms that are principally acting depends on the state of friction.

Coefficients of friction

Table 5.4 reproduces ranges of friction coefficients in various types and states of friction [5.8]. It should however be recalled that friction does not represent a constant property of a material or a combination of material properties but rather depends on the operating variables and the system structure, i. e., on the load and the elements involved in the friction process with their properties and interactions.

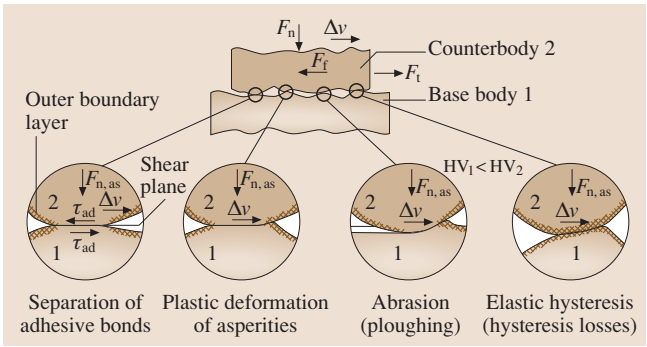


Fig. 5.7 Basic friction mechanisms viewed microscopically (F_n normal force on apparent contact area, F_f friction force between base body and counterbody, F_t tangential force, $F_{n,as}$ normal force on asperity contact, Δv relative velocity, τ_{ad} shear stress for shearing an adhesive bond, HV Vickers hardness)

5.1.3 Wear

General

As soon as the base body and counterbody come into contact, i. e., when the lubrication film thickness becomes too small or lubricant is unavailable, wear occurs. Wear is a progressive loss of material from the surface of a solid, brought about by mechanical causes, i. e., by contact and relative motion of a solid, fluid or gaseous counterbody [5.1]. Signs of wear are small detached wear particles, material removal from one friction body to the other, and material and shape changes of the tribologically loaded material zone of one or both friction partners.

Types and Mechanisms of Wear

Wear processes can be classified into different types according to the type of tribological load and the materials involved, e.g., sliding wear, fretting wear, abrasive

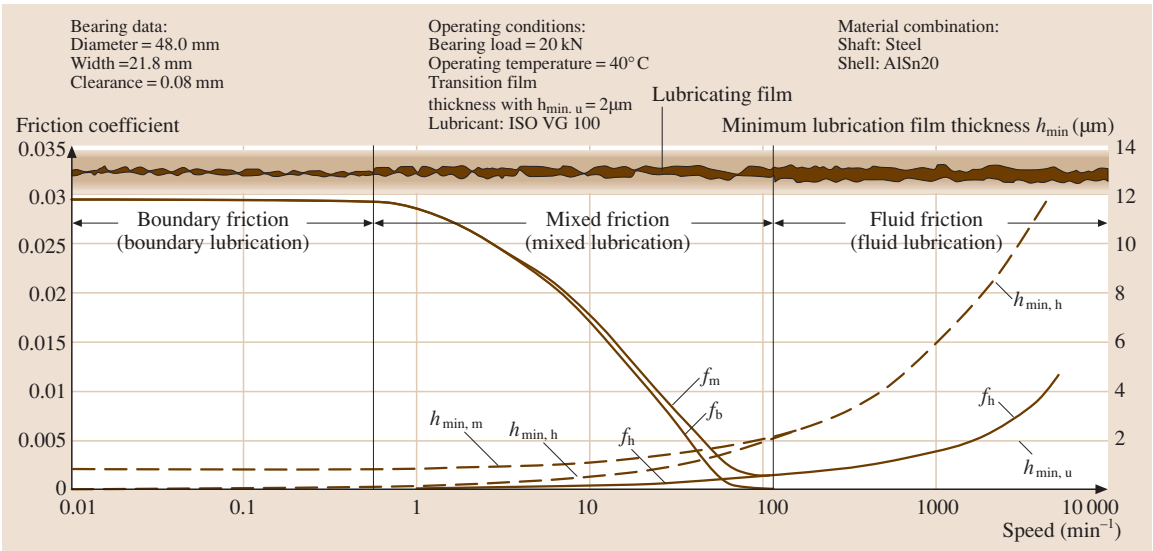
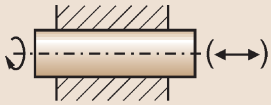
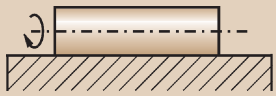
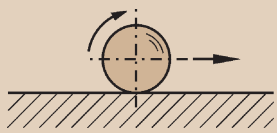
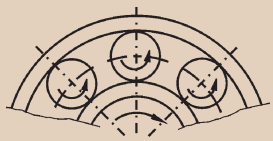


Fig. 5.6 Stribeck curve, minimum lubrication film thickness, and friction states in a radial sliding bearing (f_m friction coefficient in mixed friction, f_b friction coefficient in boundary friction, f_h friction coefficient in fluid friction, $h_{min,tr}$ minimum lubrication film thickness during the transition from fluid to mixed friction, $h_{min,tr}$ minimum lubrication film thickness during fluid friction, $h_{min,m}$ minimum lubrication film thickness during mixed friction)

Table 5.3 Contact geometry (form of contact) of tribotechnical systems

Form of contact	Base body	Counterbody	Sketch	Example application
Conformal				
Areal	Plane	Plane		Linear plain bearings
	Hollow cylinder	Solid cylinder		Sliding bearings, round fittings, cylinder races
Nonconformal				
Line contact	Plane	Cylinder		Linear roller bearings
	Cylinder	Cylinder		Roller, roller bearings
	Pinion tooth	Gear tooth		Gears
Point contact	Plane	Ball		Linear ball bearings
	Inner ring (circumferential direction)	Ball		Ball bearings

wear, and material cavitation. Wear is caused by a number of mechanisms, the following four being especially important:

- Surface fatigue
- Abrasion
- Adhesion
- Tribochemical reaction

Table 5.5 provides a breakdown of wear according to types of wear and wear mechanisms based on [5.1].

Figure 5.8 presents a chart of the effective wear mechanisms. The wear mechanisms can occur individually, successively or concomitantly.

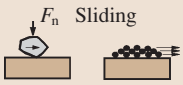
Surface fatigue manifests itself through cracking, crack growth, and detachment of wear particles, brought about by alternating loads in near-surface zones of the base body and counterbody.

In *abrasion* microcuttings, fatigue due to repeated ploughing, and fracture of the base body caused by the counterbody's hard asperities or by hard particles in the interfacial medium lead to wear.

Table 5.4 Coefficients of friction for different types and states of friction

Type of friction	Friction state	Coefficient of friction f
Sliding friction	Solid friction	0.1–1
	Boundary friction	0.1–0.2
	Mixed friction	0.01–0.1
	Fluid friction	0.001–0.01
	Gas friction	0.0001
Rolling friction	(Grease lubrication)	0.001–0.005

Table 5.5 Types and mechanisms of wear (after [5.1])

System structure	Tribological load (types of motion and simplified symbols)	Type of wear	Effective wear mechanism			
			Surface fatigue	Abrasion	Adhesion	Tribochemical reactions
Solid – Interfacial medium (complete solid seperation) – Solid	Sliding Rolling Bouncing Impacting 	–	×			×
Solid – Solid (solid friction, boundary lubrication, mixed lubrication)	Sliding 	Sliding wear	×	×	×	×
	Rolling Revolving 	Rolling wear	×	×	×	×
	Oscillating 	Fretting wear	×	×	×	×
Solid – Solid and particles	 Sliding	Sliding abrasion (three-body abrasion)		×		×
	Sliding 	Sliding abrasion (three-body abrasion)	×	×		×
	Rolling 	Rolling abrasion (three-body abrasion)	×	×		×
Solid – Fluid	Flowing Vibrating 	Material cavitation (cavitation erosion)	×			×
Solid – Fluid and particles	Flowing 	Particle erosion (erosion wear)	×	×		×

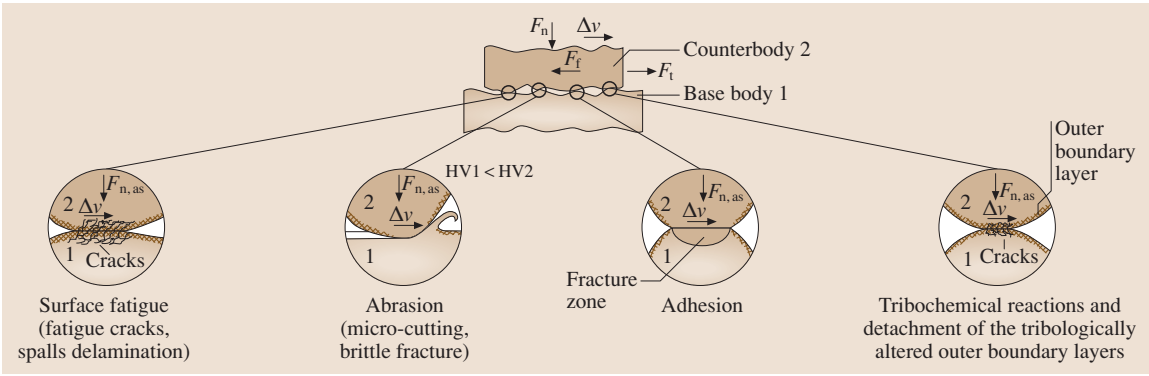


Fig. 5.8 Basic wear mechanisms viewed microscopically (F_n normal force on apparent contact surface, F_f friction force between base body and counterbody, $F_{n,as}$ normal force on asperity contact, Δv relative velocity, HV Vickers hardness)

In *adhesion*, after possibly extant protective surface layers have been broken through, atomic bonds (microwelds) form above all on the plastically deformed microcontacts between the base body and counterbody. If the strength of the adhesive bonds is greater than that of the softer friction partner, material eventually detaches from the deformed surface of the softer friction partner and is transferred to the harder one. The transferred material can either remain on the harder friction partner or detach, or even return.

In *tribochemical reactions*, friction-induced activation of loaded near-surface zones causes elements of the base body and/or counterbody to react chemically with elements of the lubricant or ambient medium. Compared with the base body and counterbody, the reaction products exhibit changed properties and, after reaching a certain thickness, can be subject to brittle chipping or even exhibit properties reducing friction and/or wear.

Apart from the types and mechanisms of wear, *wear phenomena* are also extremely interesting for interpreting the result of wear (Table 5.6). These mean the changes of a body's surface layer resulting from wear and the type and shape of the wear particles accumulating. Light or scanning electron microscope images can present this extremely clearly.

Wear Profiles and Measurable Variables

Estimating the service life of components necessitates knowing the wear profile over the loading time and/or the rate of wear (amount of wear per loading time). According to [5.1] and [5.8], different wear profiles are frequently generated depending on the effective wear mechanism (Fig. 5.9).

Three phases are distinguished: break-in, steady state, and failure. During breaking-in, increased wear, so-called break-in wear, with a degressive profile can occur and, for example, switch over into a long-lasting state with a constant increase of the amount of wear (constant rate of wear) until failure is announced by a progressive increase of wear (Fig. 5.9a).

If surface fatigue takes effect as the primary mechanism of wear, then measurable wear after break-in frequently only becomes noticeable after a certain incubation period during which microstructural changes, cracking, and crack growth commence. Wear particles only detach after the incubation period (Fig. 5.9b). A negative amount of wear is even occasionally measured at the beginning of the wear process. This is caused by material transfer from the wear partner (Fig. 5.9c).

In principle, the profile of the amount of wear W over the loading time t can be progressive, linear or degressive (Fig. 5.9d).

Table 5.6 Typical wear phenomena caused by the main wear mechanisms (after [5.1])

Wear mechanism	Wear phenomenon
Adhesion	Scuffing or galling areas, holes, plastic shearings, material transfer
Abrasion	Scratches, grooves, ripples
Surface fatigue	Cracks, pitting
Tribochemcal reactions	Reaction products (layers, particles)

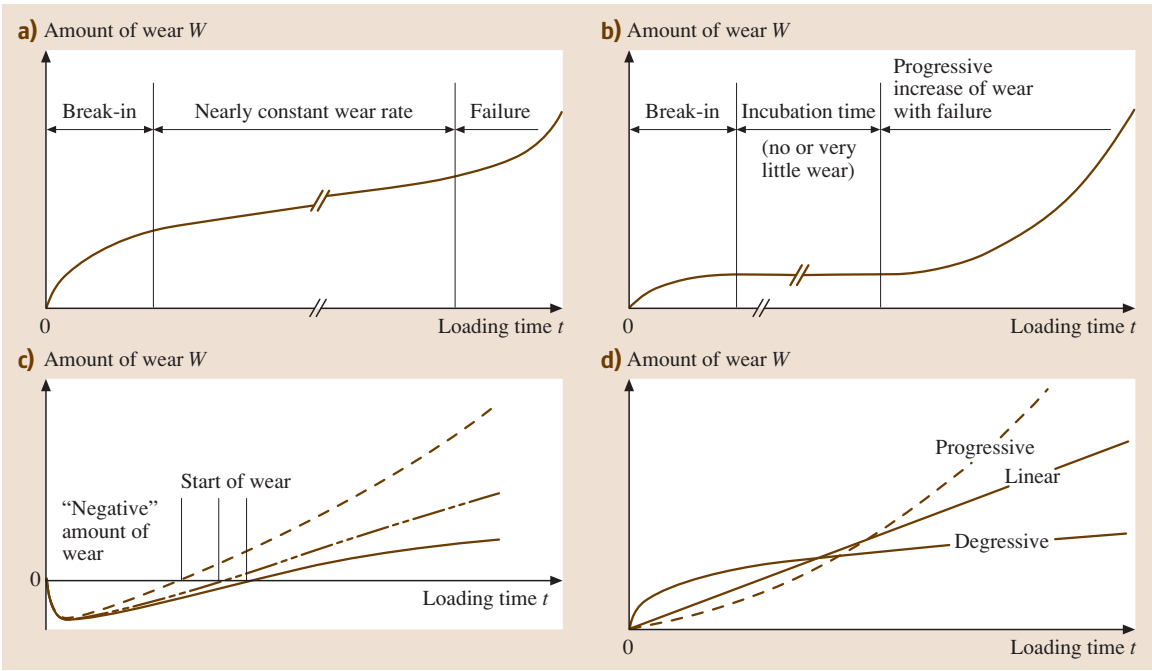


Fig. 5.9a–d Amount of wear as a function of loading time: **(a)** typical wear profile with a nearly constant wear rate; **(b)** wear profile with an incubation time; **(c)** wear profile with negative amount of wear; **(d)** different wear profiles

A difference is made between directly measurable wear variables, such as linear, planimetric, volumetric and mass amount of wear, and related measurable wear variables (wear rates), such as rate of wear, wear–distance ratio, and wear–throughput ratio. As a rule, the amount of wear can be measured. It is advisable to use specific or relative amounts of wear in comparative wear tests whenever the operating variables or the properties of the elements involved in the wear cannot be kept constant or are intentionally changed.

Determination of Wear and Useful Life

In unlubricated tribological systems and in systems operated in the boundary or mixed lubrication range, practically usable information about friction and wear can frequently only be found through experimentation since the physical and chemical mechanisms occurring here frequently cannot be described sufficiently exactly theoretically so that a precise calculation of friction and wear is often impossible.

However, mathematical equations for an approximate preliminary determination of friction and wear provide the energy balance of the friction process. If friction is approached as an energetic problem, the energy balance forms the basis for determining the friction and wear [5.9]. The friction force F_f and the friction coefficient f can be determined from the friction energy W_f

$$W_f = W_{el,\,hys} + W_{pl} + W_{abr} + W_{ad} = fF_n s_f = fF_n \Delta v t \,, \quad (5.5)$$

where $W_{el,\,hys}$ is the proportion of the work of friction from the hysteresis during elastic deformation, W_{pl} is the work of friction from plastic deformation, W_{abr} is the work of friction caused by abrasion, W_{ad} is the work of friction caused by separation of adhesive bonds, f is the friction coefficient, F_n is the normal force, s_f is the friction distance, Δv is the relative velocity between the friction bodies, and t is the friction time [5.9].

The action time of a friction pair usually depends on the wear. The anticipated useful life L_h under specific operating conditions is frequently determined by calculation. A function-based maximum allowable wear is assumed. Thus, the useful life L_h corresponds to the action time for which the friction pair can be safely operated.

Wear properties as a function of the friction time are presented in Fig. 5.10. The development of wear with its scatter as a function of the friction time is shown for various operating conditions. The height of wear h_w serves

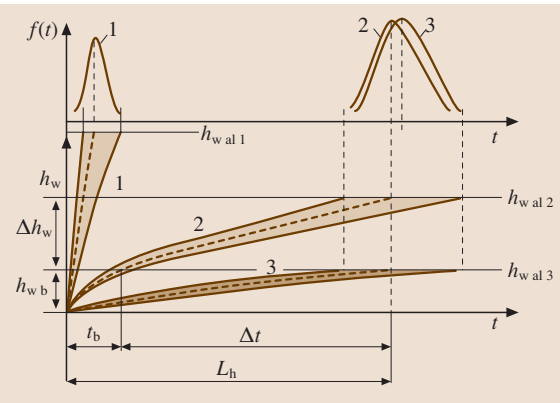


Fig. 5.10 Courses of height of wear h_w and distribution density $f(t)$ as a function of the friction time t when allowable height of wear values have been reached for differently high stresses and different frictional states (after [5.10])

as the gauge of wear. The function-based wear limit is characterized by the maximum allowable height of wear $h_{w,al}$. As a rule, a wear profile exists, which is composed of a break-in process and a steady state (Fig. 5.10, cases 2 and 3). In addition, Fig. 5.10 also shows the distribution densities $f(t)$ for the useful lives L_h defined by $h_{w,al}$.

Despite a large allowable height of wear, case 1 is not practicable since the friction pair here fails early and does not reach a steady state. Case 2 shows a normal progression in pure solid friction. More favorable conditions exist in case 3 (e.g., mixed friction), but the allowable height of wear is small.

The average useful life (action time) for the median of the distribution density results from the expression

$$L_h = t_b + \Delta t = t_b + \frac{h_{w,al} - h_{w,b}}{I_h \Delta v}\,, \tag{5.6}$$

with the linear wear intensity I_h formed from the ratio of the height of wear h_w to the friction distance $s_f (I_h = \Delta h_w / \Delta s_f)$, the break-in time t_b, and the break-in height of wear h_{wb}. If it is assumed that the wear volume V_w and the work of friction W_f are proportional ($W_f \propto V_w$), and the value e_w as a proportionality factor is introduced as the wear energy density, as a result of which the basic wear equation becomes $W_f = e_w V_w$, and if the probabilistic nature of real operation is allowed for (through the introduction of $L(\gamma)_h$ as γ percentage useful life, x as the quantile (random variable) of the standardized normal distribution, and v as an empirical coefficient of variation that represents a measure of the scatter of the wear rate), then the following expression can be set up for the useful life [5.10]

$$L(\gamma)_{h1,2} = t_{b1,2} + \frac{(h_{w,al1,2} - h_{w,b1,2}) e_{w1,2}}{\alpha_{f1,2} \tau_f \Delta v (1 - x v_{1,2})} \tag{5.7}$$

with the friction shear stress $\tau_f = f F_n / A_a$, where A_a stands for the apparent contact area, and the energy proportion factor $\alpha_{f1,2}$ is used to quantify the share of the work of friction induced in friction body 1 or 2.

For a survival probability of $\gamma = 50\%$, when there is a standardized normal distribution, the quantile is $x = 0$, whereas $x = -1.28$ applies for $\gamma = 90\%$. The coefficient of variation depends on the operating conditions (common values based on [5.11] are $v = 0.2$–0.8; when there is fatigue wear, $v = 0.2$–0.4).

Determining the useful life by using (5.7) not only requires knowledge of the break-in time t_b, the allowable height of wear $h_{w,al}$, the break-in height of wear $h_{w,b}$, the energy proportion factor α_f, the normal force F_n, the apparent contact surface A_a, the relative velocity Δv, and the statistical variables quantile x and coeffi-

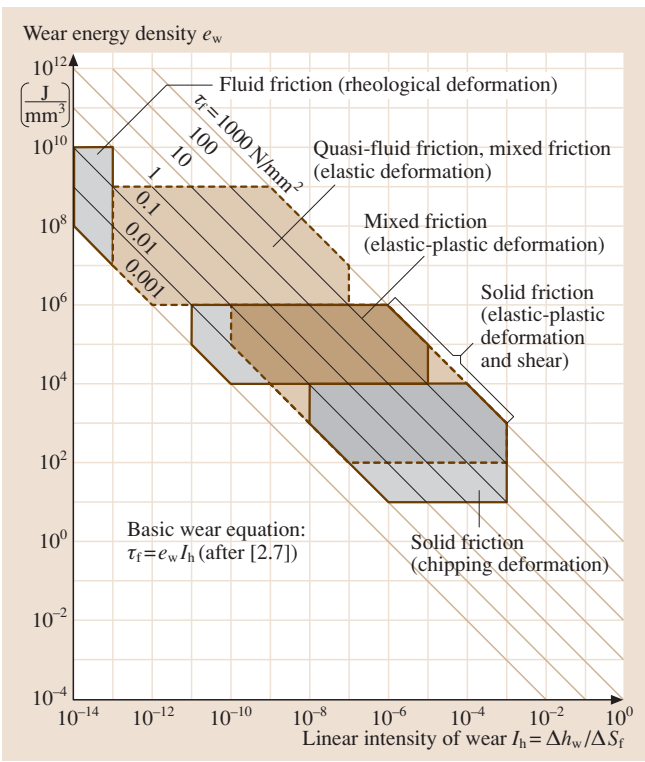
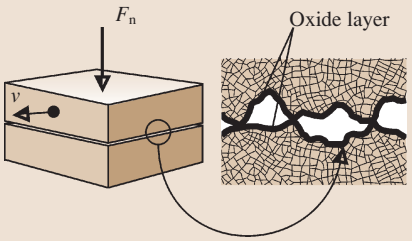
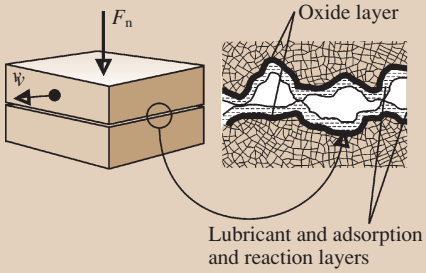
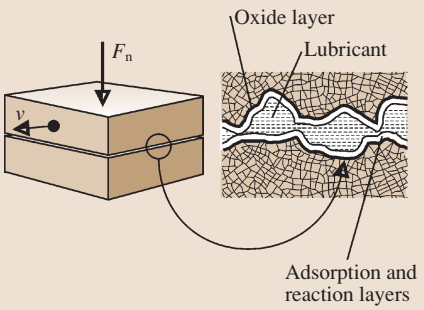
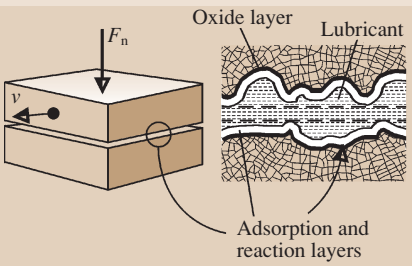


Fig. 5.11 Wear energy density e_w as a function of the linear wear intensity l_h and the friction shear stress τ_f (after [5.9])

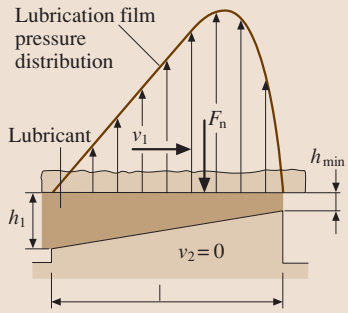
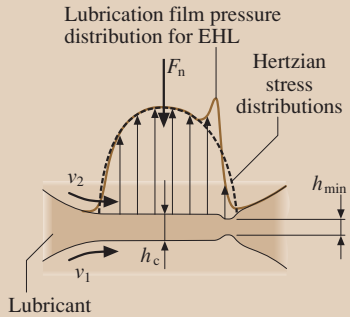
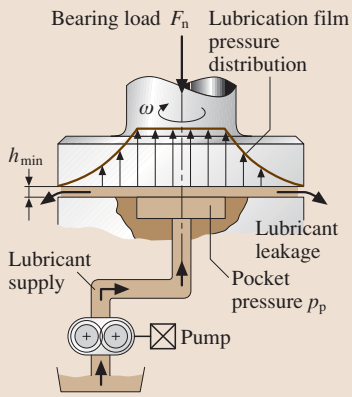
Table 5.7 Friction and lubrication states

	Friction/lubrication states
	1. Solid friction/no lubrication • Direct contact of friction partners • Formation of oxide reaction layers and adsorption of gases • High wear rates probable, risk of seizing • Friction coefficients (reference values): $f \approx 0.35$ to > 1 • Special cases *Boundary-layer friction* – Friction between solid boundary layers with modified properties compared with the bulk material (e.g., oxide layers) *Friction of pure metallic surface* – Direct contact of pure metallic surfaces (e.g., in machine or scuffing processes
	2. Boundary friction/boundary lubrication • Surfaces of the friction partners covered with a thin friction minimizing lubrication film • Physisorption, chemisorption, and tribochemical reaction with additives from the lubricant form friction-minimizing, lightly shearing layers on the surfaces • Lower wear rates than for solid friction • Friction coefficients (reference values): $f \approx 6\times10^{-2}-2\times10^{-1}$
	3. Mixed friction/mixed lubrication • Lubrication film not thick enough to seperate surfaces from each other completely; consequence: roughness contacts • Load is partially carried by lubricating film through hydrodynamic effect and partly by the roughness contacts • As in boundary friction, use of additives in lubricant is also important to generate friction-minimizing adsoption and reaction layers on the surfaces • Wear rates are smaller, the larger the hydrodynamic part of the load-carrying capacity • Friction coefficients (reference values): $f \approx 10^{-3}-10^{-1}$
	4. Fluid friction/fluid film lubrication • Friction partners are completely separated from each other by a fluid film that can be generatedhydrodynamically or hydrostatically • Virtually wear-free operation • Friction coefficients (reference values): $f \approx 6\times10^{-4}-5\times10^{-3}$

cient of variation v but above all also information about the friction coefficient f and the wear energy density e_w appearing in the contact.

The friction coefficient can be determined experimentally or estimated through calculations [5.3, 12]. Values for the wear energy density e_w are either de-

Table 5.8 Different types of fluid film lubrication

	**Hydrodynamic** [5.13] for a given geometry $h_{min} \sim \sqrt{\frac{\overline{v}\eta}{F_n}} = (\overline{v}\eta)^{0.5} F_n^{-0.5}$, $\overline{v} = \frac{v_1+v_2}{2}$ entraining velocity, η average viscosity of the lubricant in the lubricating film, F_n load
	**Elastohydrodynamics** [5.6] a) *Hard material surfaces (hard EHL)* elliptical contact surfaces, given geometry $h_{min} \approx (\overline{v}\eta_0)^{0.68} \alpha^{0.49} E^{*-0.117} F_n^{-0.073}$ $\frac{h_{min}}{h_c} \approx 0.56$ $\overline{v} = \frac{v_1+v_2}{2}$ entraining velocity, η_0 viscosity of the lubricant at gap entry at $p=0$, α viscosity-pressure coefficient, $\frac{1}{E^*} = \frac{1}{2}\left(\frac{1-\nu_1^2}{E_1} + \frac{1-\nu_2^2}{E_2}\right)$ reduced modolus of elasticity, E_1 and E_2 modulus of elasticity of friction bodies 1 and 2, ν_1 and ν_2 Poisson's ratio of friction bodies 1 and 2, F_n load. b) *Soft material surfaces (soft EHL)* elliptical contact surfaces, given geometry $h_{min} \approx (\overline{v}\eta_0)^{0.65} E^{*-0.44} F_n^{-0.21}$; $\frac{h_{min}}{h_c} \approx 0.77$
	**Hydrostatic** [5.13] for a given geometry and constant lubricant volumetric flow ($\dot{V}$ = const.) $h_{min} \approx \sqrt[3]{\frac{\eta}{F_n}} = \eta^{0.33} F_n^{-0.33}$, F_n load, η average viscosity of the lubricant in the lubricant cap

termined in tests or taken out of [5.9–11]. Figure 5.12 presents the nomogram of the basic wear equation $W_f = e_w V_w$, with ranges for typical friction and wear states for evaluating and classifying the tribological behavior of friction pairs.

5.1.4 Fundamentals of Lubrication

Lubrication completely or partially separates the surfaces of the friction bodies by selectively introducing an interfacial medium (lubricant) that minimizes friction

Table 5.9 Characteristics of various synthetic oils (after [5.15])

	Mineral oil	Polyalpha-olefine	Polyglycol (water insoluble)	Ester	Silicon oil	Alcoxi-fluorine oil
Viscosity at 40° (mm^3/s)	2–4500	15–1200	20–2000	7–4000	4–100 000	20–650
Use for oil-sump temperature (°C)	100	150	100–150	150	150–200	150–220
Use for oil-sump temperature (°C)	150	200	150–200	200	250	240
Pour point (°C)	−20 [b]	−40 [b]	−40	−60 [b]	−60 [b]	−30 [b]
Flash point (°C)	220	230–260 [b]	200–260	220–260	300 [b]	–
Evaporation losses [d]	○	+	○ to –	+	+ [b]	++ [b]
Water resistance [e]	+	+	+ [b,f]	+ to ○	+	+
η–T behavior [e]	○	+ to ○	+	+	++	+ to ○
Pressure–viscosity coefficient ($10^8\ m^2/N$) [d,c]	1.1–3.5	1.5–2.2	1.2–3.2	1.5–4.5	1.0–3.0	2.5–4.4
Suitable for high temperatures (≈ 150 °C) [e]	○	+	+ to ○ [b]	+ [b]	++	++
Suitable for high load [e]	++ [a]	++ [a]	++ [a]	+	– [b]	+
Compatibility with elastomers [e]	+	+ [b]	○ [g]	○ to –	++	+
Price ratios	1	6	4–10	4–10	40–100	200–800

[a] with EP additives; [b] dependent on type of oil; [c] measured up to 2000 bar, amount is dependent on type of oil and the viscosity; [d] very low, ++; low, +; moderate, ○; moderate to high, ○ to –; [e] exelent, ++; good, +; moderate to good, + to ○; moderate, ○; moderate to poor ○ to –; poor, –; [f] difficult to seperate since density is identical; [g] inspect when coating

and wear. Most lubricants are fluids (mineral oils, synthetic oils, water, etc.), yet they can also be solid, e.g., for use in dry sliding bearings [polytetrafluoroethylene (PTFE), graphite, molybdenum disulfide (MoS_2), etc.]. Greases are also applied, e.g., in ball bearings and sliding bearings and occasionally in gears too. Gases (air) are also employed, e.g., in gas-lubricated bearings. Figure 5.12 reproduces the areas of application of different lubricants. Powders, lubricating varnishes, and pastes can be classified as solid lubricants [5.14].

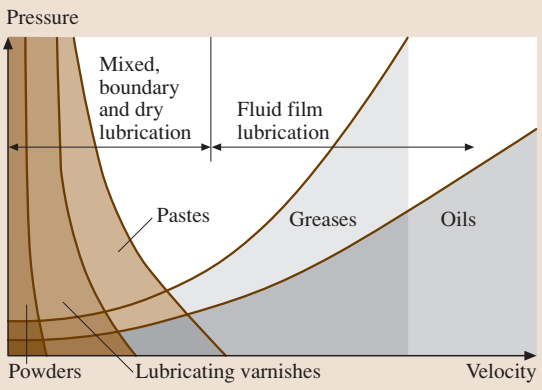


Fig. 5.12 Areas of lubricant application (after [5.14])

Similar to friction states, different lubrication states can also be defined (Table 5.7), to be precise: fluid film lubrication, mixed lubrication, boundary lubrication, and solid lubrication (lubrication with solid lubricants and surface coatings).

Fluid Film Lubrication

Fluid film lubrication, i. e., the complete separation of the frictional surfaces by a lubricating film, can be achieved by hydrodynamic, elastohydrodynamic or hydrostatic lubrication (Table 5.8). A load-dependent pressure is developed in the lubricating film, as a result of which the load can be carried.

Hydrodynamic Lubrication

The following conditions must be met to produce hydrodynamic lubrication. A viscous lubricant that adheres to both the moving and the fixed friction body must be used (adhesive effect of the lubricant). Furthermore, the friction body surfaces must be converging and the lubricant must be dragged into the converging gap. The entraining velocity depends on the velocity at which the lubricant is dragged into the contact and is frequently confused with the relative velocity. The latter is decisive for friction, while the entraining velocity is essential for load-carrying capacity.

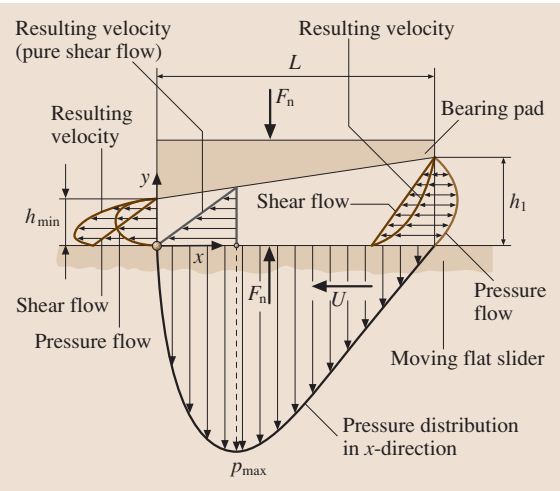


Fig. 5.13 Velocity distributions and pressure distribution in a bearing pad

The lubricant adhering to the surface is dragged into the lubrication gap by the friction bodies moving relative to the minimum lubrication film thickness. Since the gap converges in the direction of motion, the quantity of oil dragged by the friction bodies accumulates in front of the narrowest gap cross section and cannot be completely transported by the moving friction bodies' drag effect alone through the minimal cross section.

Thus, overpressure is inevitably produced in the converging zone in between the sliding surfaces, which grows to be just large enough that the difference between the quantities supplied and discharged by the drag flow are squeezed out of the lubricant gap as a result of compressive flows (satisfaction of the condition of continuity). The overpressure produces the friction contact's load-carrying capacity in equilibrium with the external load.

Figure 5.13 presents examples of velocity distributions at various points in the lubrication gap and the pressure distribution for an inclined plane bearing pad. The level of the average pressure developed (normally < 7 MPa) is usually not high enough to cause significant elastic deformation of the surfaces. The hydrodynamic pressure is dependent on the friction bodies' geometry, the inclination of the surfaces to each other, the viscosity of the lubricant, the load, and the velocity of the lubricant dragged into the converging lubrication gap.

Table 5.8 specifies the correlation between the minimum lubrication film thickness h_{min} and the entraining velocity, the viscosity of the lubricant, and the load. It is evident that the minimum lubrication film thickness increases as the entraining velocity increases and the viscosity becomes greater, and that it decreases as the load grows, although only with an exponent of 0.5 in each case. The minimum lubrication film thickness is normally greater than 1 μm.

Above all, it is the lubricant's viscosity that determines the lubricating effect in hydrodynamic lubrication. Friction is only generated by the shear of the viscous lubricant. Apart from the viscosity, the relative velocity between the moving bodies' lubrication gap surfaces and the lubrication gap height are decisive for friction. If both surfaces slide with the same sliding velocity, the relative velocity is $\Delta v = 0$, i. e., friction does not occur. However, when the sliding conditions are identical, the entraining velocity in the lubrication gap can be high (e.g., in the case of radial sliding bearings, in which the shaft and bushing rotate with the same speed in the same direction) so that the tribotechnical system then has a high load-carrying capacity and a friction force of zero.

Elastohydrodynamic Lubrication (EHL). Elastohydrodynamic lubrication (EHL) is a form of hydrodynamic lubrication in which elastic deformations of the lubricated surfaces become significant. The essential prerequisites for hydrodynamic lubrication, such as a converging lubricating film, the lubricant entraining velocity in the converging gap, and a viscous lubricant between the surfaces, are also important in EHL. Elastohydrodynamic lubrication is normally connected with nonconformal surfaces (Table 5.3). There are two different forms of EHL: for hard surfaces (hard EHL) and soft surfaces (soft EHL).

Hard EHL. Elastohydrodynamic lubrication for hard surfaces (hard EHL) refers to materials with high moduli of elasticity, e.g., metals. Both the elastic deformations and the viscosity's pressure dependence are equally important in this type of lubrication. The maximum occurring lubricating film pressure is typically between 0.5 and 4 GPa. The minimum lubrication film thickness normally exceeds 0.1 μm. When the loads are those normally occurring in nonconformal contacts of machine elements, the elastic deformations in hard EHL exhibit values several magnitudes greater than the minimum lubrication film thicknesses. Furthermore, the lubricant viscosity in the lubrication film can change by a magnitude of 3–4 or more, depending on the lubricant, pressure, and temperature.

The minimum lubrication film thickness h_{min} is a function of the same parameters as in hydrodynamic

lubrication, but these must be augmented by the effective modulus of elasticity E^* and the lubricant's viscosity–pressure coefficients α. Table 5.8 indicates that, in the relationship for the minimum lubrication film thickness, the exponent for the normal load in hard EHL is approximately seven times smaller than it is in hydrodynamic lubrication. This means that, in contrast to hydrodynamic lubrication, the load only marginally influences the lubrication film thickness in hard EHL. The reasons are to be found in the increase of the contact area as the load increases in hard EHL, as a result of which a larger lubrication area is provided to bear the load. The exponent for the lubricant entraining velocity in hard EHL is greater than in hydrodynamic lubrication. Typical applications for hard EHL are toothed gears, rolling element bearings, and cam–follower pairs.

Soft EHL. Elastohydrodynamics for soft surfaces (soft EHL) refers to materials with low moduli of elasticity, e.g., rubber. In soft EHL, sizeable elastic deformations occur even at low loads. The maximum occurring pressures in soft EHL are typically 1 MPa, in contrast to 1 GPa for hard EHL. This low lubricating film pressure only negligibly influences the viscosity during the flow through the lubrication gap. The minimum lubrication film thickness is a function of the same parameters as in hydrodynamic lubrication with the addition of the effective modulus of elasticity E^*. The minimum lubricating film thickness in soft EHL is typically 1 μm. Applications for soft EHL are seals, artificial human joints, tires and nonconformal contacts in which rubber is used.

A common feature that hard and soft EHL exhibit is the generation of coherent lubricating films as a result of local elastic deformations of the friction bodies and thus the prevention of interactions between asperities. Hence, only the lubricant's shear generates frictional resistance to motion.

Hydrostatic Lubrication. In hydrostatic lubrication of friction bodies, a pocket or recess is incorporated in one friction body's loaded surface into which a fluid is forced from outside at constant pressure. A pump outside the bearing generates the lubricant pressure. Hence, the lubricant pump and the lubricating pocket into which the lubricant is fed under pressure are the most important features of hydrostatic lubrication. The lubricating pocket is normally positioned opposite the external load. The load-carrying capacity of a contact with hydrostatic lubrication is also assured when surfaces are not moving. When the volumetric flow of lubricant into the lubricating pocket is constant, the minimum lubrication film thickness is proportional to the cube root of the ratio of the average lubricant viscosity in the lubrication gap and the load, i. e., the minimum lubrication film thickness is less dependent on the viscosity and the load than is the case in hydrodynamic lubrication.

Hydrostatic lubrication is mainly used: where the friction partners' surfaces do not have any metallic contact, i. e., wear may not occur, not even when ramping up and ramping down a machine or at low speed; where as low a friction coefficient as possible must be produced at low speeds; and where, as a result of less effective lubricant entraining velocities in the lubrication gap, the wedge effect cannot produce any bearing lubricating film hydrodynamically.

Boundary Lubrication

In *boundary lubrication*, the friction bodies are not separated by a lubricant, the hydrodynamic lubricating film effects are negligible, and there are extensive asperity contacts. The physical and chemical properties of thin surface films of molecular thickness control the lubricating mechanisms in the contact. The base lubricant's properties are of little importance. The coefficient of friction is on the whole independent of the lubricant's viscosity. The frictional characteristic is determined by the properties of the solids involved in the friction process and the boundary layers forming on the material surfaces, which primarily depend on the lubricant's properties, particularly the lubricant additives, as well as the material surfaces' properties. These boundary layers are formed by physisorption, chemisorption, and/or tribochemical reaction. The thickness of the surface boundary layer varies between 1 and 10 nm, depending on the molecule size.

In *physisorption*, additives contained in the lubricant [e.g., antiwear (AW) additives] such as saturated and unsaturated fatty acids, natural and synthetic fatty acid esters, and primary and secondary alcohols are adsorbed on the tribologically loaded surfaces. Such materials have in common a high dipole moment because of at least one polar group in the molecule (Fig. 5.14).

The coverage of the surfaces follows the laws of adsorption and is dependent on temperature and concentration. A prerequisite for the adsorption of polar groups is that the material surface exhibits a polar character so that van der Waals bonds can form. This is usually attained for metallic materials by oxide films

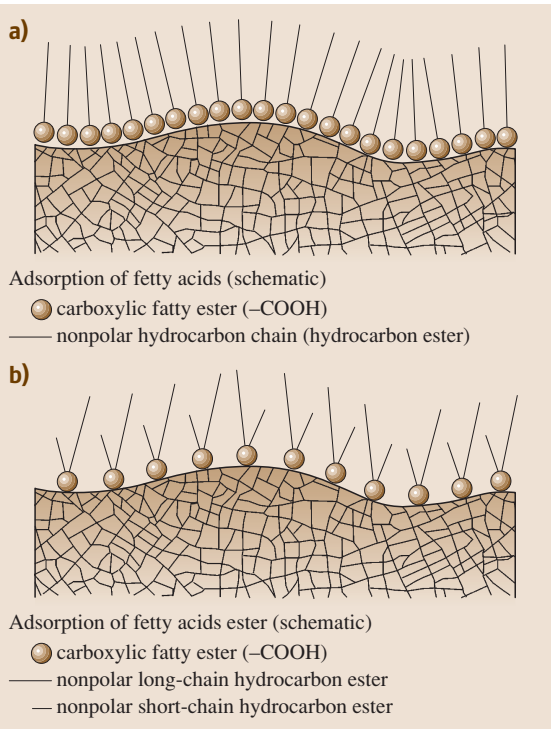


Fig. 5.14a,b Adsorption on metal surfaces (after [5.16])

forming on the surfaces. Problems can arise for ceramic materials however. While, for example, adsorption of fatty acids with polar end groups occurs easily on aluminum oxide with ionic bonding so that friction is diminished from a certain chain length onward, adsorption apparently does not occur on silicon carbide with covalent bonds so that the friction coefficient is not influenced [5.2].

In *chemisorption*, molecules are bonded to the surface. Substantially more-stable boundary layers develop because chemical bonds with greater bonding force are formed on the surface (e.g., the reaction of stearic acid with iron oxide when water is present, as a result of which metallic soap forms as iron stearate). Chemisorbed layers have excellent lubricating properties up to their melting point. At medium loads, temperatures, and velocities, they produce a sustained reduction of friction.

Tribochemical reactions between elements of the lubricant and the metallic material surface form reaction layers that generally have more thermal and mechanical load-carrying capacity than layers formed by physisorption or chemisorption. For this, chlorine, phosphorous or sulfur compounds are added to the lubricants as

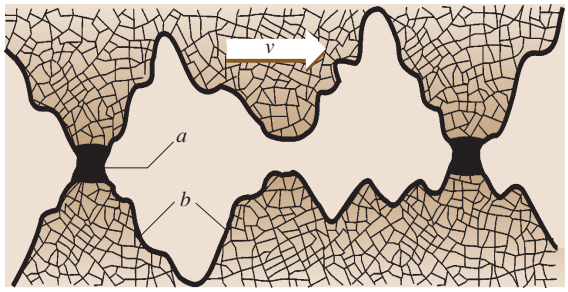


Fig. 5.15 Formation of a protective layer by EP active ingredients of sulfur compounds. a: metal–sulfide compound; b: metal surface (after [5.16])

additives [extreme-pressure (EP) additives]. The effectiveness of such EP additives depends on the speed of reaction layer formation, which is influenced by the reaction's activation energy, the surface temperature, and the concentration of the additive. EP active ingredients themselves, or their thermal cleavage products, react with the surface metallic oxide in very short time (10^{-6}–10^{-7} s), forming a reaction layer that adheres well and shears easily (Fig. 5.15).

While only very little or no wear appears in hydrodynamic and elastohydrodynamic lubrication because there are no asperity contacts, the number of asperity contacts grows in boundary lubrication and thus so does the rate of wear as the load increases. However, compared with unlubricated conditions, the rates of wear in boundary lubrication are substantially lower.

Partial or Mixed Lubrication

If the load in hydrodynamically or elastohydrodynamically lubricated machine elements grows too large or the lubricant entraining velocity in the lubrication gap is too small, the lubricating film becomes too thin so that asperity contacts appear at some points. This is mixed friction, in which the states of boundary friction and fluid friction coexist. The lubricating film is no longer coherent because individual surface asperities of the paired parts penetrate it and cause direct contact of the sliding surfaces. In partial or mixed lubrication, the normal load is partly carried by hydrodynamic film pressures and partly by contact of the frictional surfaces at the asperity contacts. The transition from hydrodynamic or elastohydrodynamic lubrication to partial lubrication does not happen suddenly when the load is increased but rather the share of the load carried by the hydrodynamic pressure buildup decreases gradually, while the share carried by the asperities grows in equal measure.

Fig. 5.16 Solid lubrication made of lubricating varnish: a: phosphate layer (3 µm); b: binder; c: solid lubricant; d: base metal (after [5.16])

Lubrication with Solid Lubricants and Surface Coatings

In lubrication with solid lubricants and in coatings, the most important measures for improving friction, preventing scuffing or seizing, and minimizing wear are applying metallic (e.g., indium) or nonmetallic layers (e.g., synthetic resin coatings), forming reaction layers through chemical transformation (e.g., phosphatizing), and using solid lubricants. Nonmetallic layers also include layers obtained by physical vapor deposition (PVD) or chemical vapor deposition (CVD) from the gas phase; the layer has a thickness in the micron range and can be made of, for example, titanium nitride (TiN), titanium aluminum nitride (TiAlN), chromium nitride (CrN) or tungsten carbide/carbon (WC/C).

Among the solid lubricants, crystalline solids with layer lattices (lamina structure) appear particularly suitable. Their structure is characterized by lamellae that can be displaced slightly with respect to each other along glide planes. Typical representatives are graphite (C) and molybdenum disulfide (MoS_2). Certain plastics such as PTFE are also used as solid lubricants.

Apart from these, coating layers on metals, which are adhesive, act as lubricants, and are firm to the touch, so-called lubricating varnishes, also play a role (Fig. 5.16). They have a large proportion of solid lubricants and can be used as dry films over a wide range of temperatures.

5.1.5 Lubricants

Gaseous, fluid, consistent or solid lubricants are used in mechanical engineering. Gaseous lubricants are utilized, for example, in high-speed, lightly loaded machines (e.g., ultracentrifuges, gas pumps for nuclear power plants) or in equipment technology. The most important group of substances for fluid lubricants are the mineral oils, as well as synthetic, animal, and vegetable oils and even water in special applications. Greases and adhesive lubricants are consistent lubricants. Greases consist of mineral or synthesis oils that are thickened with, for instance, soaps. There are bitumen-based adhesive lubricants as well as bitumen-free adhesive spray lubricants. These are principally used to lubricate larger, open-toothed gears. Solid lubricants are frequently introduced into fluid or consistent carrier substances. Solid lubricants are used in pure form only under special operating conditions. Among others substances, solid lubricants include graphite, molybdenum disulfide, PTFE, etc.

During operation, pressure and shear stresses, among others, act on a lubricant at different temperatures. The lubricant comes into contact with gases (e.g., air), liquids (e.g., water), and solids (e.g., metals, sealing materials, wear particles). In addition, contact with human skin cannot be ruled out. This results in a number of requirements that a good lubricant should have. For a particular application, it should have an appropriate viscosity–temperature and viscosity–pressure behavior, a low pour point, and preferably not be volatile, exhibit high-temperature, shear, and oxidation stability, and be hydrolysis and radiation resistant in special cases. Beyond this, a good lubricant should be compatible with the constructional materials used and be nontoxic and not cause any disposal problems.

Lubricating Oils

Among the fluid lubricants, a difference is made between mineral oils and synthetic oils.

Mineral Oils. Mineral oils are obtained from naturally deposited crude oil with the aid of distillation and refining. Crude oils chiefly consist of hydrocarbons and organic oxygen, sulfur, and nitrogen compounds. The exact composition of a crude oil depends on its origin (provenance). When mineral oils are manufactured, systematic selection of the base oil and control of the manufacturing process can produce particular compositions and thus influence desired properties. Mineral oils are mixtures of hydrocarbons, which can be subdivided as a function of their structure into open-chain hydrocarbons (paraffins) and cyclic hydrocarbons (naphthenes, aromatics), which can each be saturated or unsaturated.

Mineral oils' performance characteristics change as their working life increases. A primary reason for diminished performance is gradual oxidization of hydrocarbon chains, with aromatics reacting more than naphthenes and naphthenes more than paraffins. As a result, sludge-like deposits can form, which impede the oil supply by clogging oil feed lines and filters. Moreover, the formation of organic acids can be boosted,

Table 5.10 Comparison of the properties of natural and synthetic base oils for lubricants (after [5.16])

	A	B	C	D	E	F	G	H	I	J	K
Viscosity–temperature behavior (viscosity index, VI)	–	+	○	+	+	○	– –	– –	++	++	–
Low-temperature performance (pour point)	– –	○	++	++	+	○	– –	–	++	+	○
Oxidation stability (aging test)	–	– –	+	○	+	– –	+	○	+	+	++
Thermal stability (heating under absence of oxygen)	–	–	–	○	+	○	++	–	+	○	+
Volatility (evaporating loss)	–	○	○	++	++	○	○	+	+	○	○
Finish compatibility (effect on coatings)	++	–	++	–	–	–	–	– –	○	–	○
Water resistance (hydrolysis test)	++	– –	++	–	–	○	++	–	○	–	+
Antirust properties (corrosion test)	++	++	++	–	–	○	–	–	○	– –	+
Seal compatibility (swelling behavior)	○	–	++	–	–	○	○	– –	○	○	–
Flame resistance (ignition temperature)	– –	– –	– –	–	–	–	–	++	○	–	++
Additive solubility (dissolving of larger concentrations)	++	○	+	– –	○	–	+	○	– –	○	–
Lubricity (load-carrying ability)	○	++	○	+	+	○	++	++	– –	–	++
Biodegradability (degradability test)	–	++	○	++	++	++	– –	+	– –	–	– –
Toxicity	○	++	++	○	○	+	○	– –	++	–	+
Miscibility (formation of a homogenous phase)	++	++	++	+	+	– –	○	–	– –	–	– –
Price ratio to mineral oil	1	3	4	7	8	8	350	7	65	25	350

Weighting: 1: ++; 2: +; 3: ○; 4: –; 5: – –

A – Mineral oil (solvent neutral); B – Rape oil; C – Polyalphaolefin; D – Carboxylic acid ester; E – Neopentyl polyol esters; F – Polyalkylenglycol (polyglycol); G – Polyphenyl ether; H – Phosphoric acid ester; I – Silicon oil; J – Silicate ester; K – Fluorine-chlorine-carbon oil (chlorotrifluoroethylene)

which can cause corrosion of machine parts. This can be prevented in part by admixing additives (e.g., antioxidants, detergent, and dispersant agents). More information on their effect and the use of additives can be found in the section on "Additives."

Synthetic Oils. Synthetic-base lubricating oils are produced by chemical synthesis from chemically defined structural elements (e.g., ethylene). Their development has made it possible to systematically satisfy even extreme requirements (e.g., lubricant temperature $> 150\,°C$). According to their chemical composition, synthetic lubricants are subdivided into synthetic hydrocarbons, which only contain carbon and hydrogen [e.g., polyalphaolefines (PAO), dialkylbenzenes (DAB), polyisobutenes (PIB)], and synthetic fluids (e.g., polyglycols, carboxylic acid esters, phosphoric acid esters, silicon oils, polyphenyl ethers, fluorine–chlorine–carbon oils). Typical characteristics of synthetic oils are provided in Table 5.9 and a comparison of the properties of synthesis oils with those of mineral oil is presented in Table 5.10.

Synthetic oils have a number of advantages over mineral oils. They have better resistance to aging (thermal and oxidative stability) and thus their useful life is three to five times longer. They exhibit a more favorable viscosity–temperature behavior (with a significantly lower dependence of viscosity on temperature), display better flow properties at low temperatures and lower volatility at high temperatures, can cover applications operating at a substantially expanded range of temperature, and are radiation and flame resistant. Moreover, synthetic lubricants can be used to obtain specific frictional properties, e.g., lower friction coeffi-

Table 5.11 Examples of use of the most important synthetic lubricants (after [5.17])

Product group	Examples of use
Polyalphaolefins (synthetic hydrocarbons)	– High-performance oils for diesel engines – Multigrade engine oils – Gear lubrication at high thermal stress – Compressor oils
Carboxylic acid esters	– Aircraft engine oils – Fuel economy oils (low-friction engine oils) – Base oil for high- and low-temperature greases – Applications requiring good and fast biodegradability
Neopentyl polyol esters	– Applications similar to those for carboxylic acid esters but especially wherever oxidation stability and better additive solubility are required
Polyalkylglycols (polyglycols)	– Metalworking fluids – Gear oils (worm gears) – Hydraulic fluids (flame resistant) – Lubricant for compressors and pumps
Polyphenyl ethers	– High-temperature lubricants (up to 400 °C) – Applications requiring resistance to ionizing radiation (γ rays and thermal neutrons)
Phosphoric acid esters	– Plasticizers – Flame-resistant hydraulic oils – Safety lubricants for air and gas compressors – EP additives
Silicone oils	– Special lubricants for high temperatures – Base oil for lifetime lubricating greases (e.g., for clutch release bearings for motor vehicle clutches, starters, brakes, and axle components)
Silicate esters	– Hydraulic oils for lower temperatures – Heat exchange fluids
Fluorine-chlorine-carbon oils	– Lubricants for oxygen compressors and for pumps for aggressive fluids

cients to minimize power loss in ball bearings or gears, or higher friction coefficients to increase the transmittable torque in friction gears.

On the other hand, synthetic lubricants often cannot be used as universally as mineral oils since they have been developed for specific properties. In addition, they are more strongly hydroscopic (water attracting), display only slight air release characteristics (risk of foaming), mix poorly or not at all with mineral oils, are toxic to a large extent, and are characterized by poor compatibility with other materials (risk of chemical reaction with seals, paints, and nonferrous metals) and by poor solubility for additives. They are not always available, most notably in certain viscosity classes, and they frequently cost substantially more. Table 5.11 details examples of typical areas of application of synthetic oils.

Biodegradable Oils. Environmentally compatible lubricating oils are increasingly being used, for example, in motor vehicles and equipment in water protection areas and in hydraulic engineering, in vehicles for agriculture and forestry, and in openly running gears with loss lubrication (excavators, mills). They are readily and rapidly degradable, have a low water hazard class, and are toxicologically harmless. Their base substances have to be degraded in a degradability test (e.g., CEC L-33-T-82) by a defined amount within a specified time and the additives used (up to a maximum of 5%) should be potentially degradable. Native oils and native base synthetic esters as well as fully synthetic esters and polyglycols are used. Native oils (e.g., rape oil and natural esters) are unsuitable for high temperatures (> 70 °C) and additionally have low thermal stability and resistance to aging. The synthetic oils suitable for continuous high temperatures are often used as hydraulic oils in agricultural and forestry machines. Polyglycols are used, for example, as readily biodegradable oils in water engineering.

Table 5.12 Additives, typical types of additives, applications, and active mechanisms (after [5.18])

Additive	Types of additive	Application	Active mechanisms
Antiwear (AW) additive	Zinc dialcyldithiophosphates, tricresylphosphates	Decrease of inordinate wear metal surfaces	Reaction with metal surfaces produces layers that are plastically deformed and improves the contact pattern
Extreme pressure (EP) additives	Sulfurized greases and olefines, chlorohydrocarbons, lead salts of organic acids, aminophosphates	Prevention of microwelding between metal surfaces at high pressures and temperatures	Reaction with metal surfaces produces new bonds with lower shear resistance than the base metal. There is constant shearing off and reformation
Friction modifiers	Fatty acids, fat amines, solid lubricants	Reduction of friction between metal surfaces	Highly polar molecules are absorbed on metal surfaces and seperate the surfaces, solid lubricants form friction-reducing surface film
Viscosiy index improvers	Polyisobutylenes, polymethylacrylates, polyacrylates, ethylenepropylene, styrene maleic acid esters, copolymers, hydrogenated styrene-butadiene-copolymers	Reduction of dependence of viscosity on temperture	Polymer molecules are strongly balled in cold oil (poor solvent) and take on greater volume in warm oil (good solvent) by unballing. This produces a relative thickening in oil
Pour point depressants	Paraffin-alkylated naphthalenes and phenols, polymethylacrylates	Decrease of pour point of the oil	Encasing prevents the agglomeration of paraffin crystals
Detergent additives	Normal or alkaline calcium, barium or magnesium-sulfinates, phenates or phosphonates	Reduction or prevention of deposits in engines at high operating temperatures	Reaction with the oxidation products controls the formation of coating and sludge. Products are produced that are oil soluble or suspendet in oil
Dispersant additives	Polymers such as nitrogenous polymethylacrylates, alkyl succunimides and succinate esters, high molecular weight amines and amides	Prevention or delay of the development and deposition of sludge at low operation temperatures	Dispersantshave a pronounced affinity for impurities and encase these with oil soluble molecules that suppress the agglomeration and deposition of sludge in the engine
Oxidation inhibitors	Inhibited phenols, amines, organic sulfides, zinc dithiophosphates	Minimization of the formation of resin, coating, sludge, acid, and polymer-like compounds	Reducing the organic peroxides ends the oxidation chain reaction. reduced oxygen intake by the oil decreases the acid formation. Catalytic reactions are prevented
Corrosion inhibitors	Zinc dithiophosphates, sulfurized terpenes, phosphorized, sulfurized olefines	Protection of bearing and other metal surfaces against corrosion	Acts as an anticatalyst; film forms on metal surfaces as protection against attacks from acids and peroxides
Rust inhibitors	Amine phosphates, sodium, calcium, and magnesium sulfates, alkyl succinic acid, fatty acids	Protection of ferrous surfaces against rust	Metal surfaces prefer to adsorb polar molecules and they serve as a barrier against water neutralization by acids

Additives. Additives are substances that either give new characteristics to mineral, synthesis or vegetable oils or enhance already existing positive properties. The quantity of additive used differs greatly. Thus, circulating or hydraulic oils may only contain 0.1%, whereas special engine and gear oils may contain up to 30% additives.

All properties of lubricants cannot be changed by additives. However, using additives a clear improvement in lubrication can be obtained by modifying some properties. Thus, for example, heat dissipation, viscosity–density properties, and temperature resistance cannot be influenced by additives. Improvements brought about by additives are obtained for low-temperature performance, aging stability, viscosity–temperature properties, and corrosion protection. Only additives can attain good cleaning performance, favorable dispersion behavior, antiseizing properties, and foam inhibition.

Additives have to be matched to the base oil in terms of quantity and composition and the presence of other additives since they respond differently to the base oil and are not mutually compatible in every case. For example, there are antagonistic effects between viscosity index improves and antifoam additives, between detergent/dispersant additives and antiwear, antiseizing, and

Table 5.12 (cont.)

Additive	Types of additive	Application	Active mechanisms
Metal deactivators	Triarylphosphate, sulfur compounds, diamines, dimercaptothiadiazop deriviatives	Suppression of the catalytic influence on oxidation and corrosion	A protective film is adsorbed an metal surfaces, which inhibits the contact between the bases metal and the corrosive substance
Foam inhibitors	Silicon polymers, tributylphosphates	Protection of the development of stable foam	Attacking the oil film surrounding every air bubble reduces the boundary surface stress. As a result smaller bubbles coalesce into larger bubbles that rise to the surface
Adhesion improvers	Soaps, polyisobutylenes and polyacrylate polymers	Increase of the oils's adhesive ability	Viscosity is increased. Additives are viscous and sticky
Emulsifier	Sodium salts of sulfonic acids and other organic acids, fat amine salts	Emulsification of oil in water	Adsorbing the emulsifier in the oil/water boundary surface reduces boundary surface stress, as a result of which one fluid disperse into another
Demulsifier	Anionic sulfon acid compounds (dinonylnaphthalinsulfonat)	Demulsification of water	A boundary layer develops between water and oil form substances active in the boundary surface
Bactericide	Phenols, chlorine compounds, formaldehyde derivatives	Increase of the emulsion's working life, prevention of unpleasent odors	The growth of microorganisms is prevented or delayed

antifoam additives, and between corrosion inhibitors and antiwear and antiseizing additives [5.16].

A difference can be made between additives that form surface layers and those that change the properties of the lubricant itself. Additives forming surface layers act as a lubricating film above all when there is insufficient lubrication, as a result of which friction is reduced and the load-carrying capacity of sliding–rolling pairs is improved. Among others, this group of additives includes antiwear (AW) additives, extreme pressure (EP) additives, and friction modifiers. Adding additives that form surface layers also has drawbacks though. Thus, lubricants with additives oxidize faster than normal mineral oils and corrosive acids and insoluble residues frequently form. Hence these additives should only be used when necessitated by the operating conditions. Additives that modify lubricants influence, for example, foaming behavior, corrosion behavior, sludging, and pour point. Table 5.12 provides an overview of the most important types of additives and their applications.

During operation, the effectiveness of some additives can decrease (exhaustion) since reaction with the materials or the atmospheric oxygen causes their concentration to drop. Once the concentration of the additive falls below a certain value, an oil change is necessary.

Consistent Lubricants (Lubricating Greases)

Consistent lubricants have a flow limit. No movement occurs below a shear stress that is specific to the lubricant. Only when this flow limit has been exceeded does the viscosity drop from a virtually infinitely high to a measurable value.

Lubricating greases consist of three components: a base oil (75–96 wt %), a thickener (4–20 wt %), and additives (0–5 wt %). Suitable thickeners can be dispersed both in mineral oils and in synthetic or vegetable oils so that consistent lubricants are produced. By far, most greases are manufactured using soaps (metallic salts from fatty acids) as thickeners. Thus, fatty acids are dissolved in the base oil at relatively high temperatures and a suitable metal hydroxide (e.g., hydroxides of sodium, lithium, and calcium or to a lesser extent barium and aluminum) is added subsequently. Long-chain fatty acids come from vegetable or animal oils and can be hydrogenated. Occasionally, not only long-chain fatty acids but also short-chain acids such as acetic, propionic, benzoic acid, etc. are used. Then so-called complex soaps are produced [5.16]. Most soap compounds form a fibrous matrix of interlocking particles, which retains the base oil (Fig. 5.17).

By contrast, aluminum soaps contain a spherical gel structure. The grease's lubricating action is based on the base oil being dispensed slowly and sufficiently in operation under load. The delivery of the base oil depends strongly on the temperature. The lubricating grease releases less and less oil as the temperature drops and the grease becomes stiffer and stiffer (consistency). Beyond a certain temperature limit, this eventually leads to insufficient lubrication in the friction contact. As the

Fig. 5.17 Fiber structure of a grease with soap thickener [5.19]

temperature increases, more and more oil is released. Simultaneously, the grease ages and oxidizes faster and the deterioration products produced have an adverse effect on the lubrication. A standard value is to cut the grease's working life and thus the relubrication interval in half for every 15 °C rise in temperature above approximately 70 °C. Below 70 °C, the grease's working life and consequently the relubrication interval can be extended unless the temperature drops below the lower limit. The type of base oil, its viscosity, and the additives it contains are decisive for the lubricating properties.

Lubricating greases are predominantly used at low speeds since the lubricant results in lower frictional heat transportation than oil lubrication. The relevant temperature ranges are generally between −70 °C to +350 °C. Lubricating greases also frequently have the job of protecting lubricating points from the infiltration of water and dirt and keeping out small quantities of dirt without disrupting the function. Tables 5.13 and 5.14 present performance the characteristics and applications of different lubricating greases.

Additives mainly serve to improve particular performance characteristics of the base oils. They must be uniformly distributed and be dissolved. Additives can improve the following properties of greases in particular: oxidation stability, corrosion protection, water resistance, adhesiveness, and antiwear properties.

The greatest care is necessary when mixing different types of lubricating greases since not all types of lubricating greases are mutually compatible (Table 5.15). Thus, for example, sodium soap grease is incompatible with nearly all other lubricating greases with the exception of barium complex soap grease. Lithium soap grease is incompatible with sodium soap grease, aluminum complex soap grease, and bentonite grease. In

Table 5.13 Performance characteristics of mineral-oil-based lubricating greases [5.16]

	Sodium	Lithium	Calcium	Calcium complex	Bentonite
Thickener form	Fibre	Fibre	Fibre	Fibre	Platelets
Fiber length (μm)	100	25	1	1	0.5
Fiber diameter (μm)	1	0.2	0.1	0.1	0.1
Short description	Long fibered	Medium fibered	Short fibered	Short fibered	Short fibered
Properties					
Drop point (°C)	150–200	170–220	80–100	250–300	rd. 300
Operating temperature					
Upper (°C)	+100	+130	+50	+130	+150
Lower (°C)	−20	−20	−20	−20	−20
Water resistance	Not stable	Good	Very good	Very good	good
Mechanical stability [1)] (0.1 mm)	60–100	30–60	30–60	< 30	30–60
Corrosion protection [2)]	Good	Very poor	Poor	Poor	Good
Use					
Suitability for ball bearings	Good	Very good	Variable	Variable	Very good
Suitability for sliding bearings	Good	Good	Variable	–	–
Primary use	Low-viscosity gear grease	Multipurpose grease	–	Multipurpose grease	High-temperature grease
Price	Medium	High	Low	Very high	Very high

[1)] Difference in penetration after 60 and 100 000 double strokes

[2)] Can be noticeably improved by additives

Table 5.14 Areas of application of synthetic lubricating greases [5.16]

	Mineral oil (benchmark)	PAO	Ester oils	Silicon oils	Alkoxyfluorine oils
Upper limit of application (°C)	150	200	200	250	250
Lower limit of application (°C)	−40	−70	−70	−75	−30
Lubrication of metals	++	++	+++	– – –	–
Lubrication of plastics	○	++	○	+++	+++
Hydrolysis resistance	++	++	○	+++	+++
Chemical resistance	+	+	– –	++	+++
Elastomer compatibility	○	+	○	+++	+++
Toxicity	–	+	+	+++	+++
Flammability	– – –	– – –	+	++	+++
Radiation resistance	– –	– –	–	+	++

+++ excellent; ++ very good; + good; ○ moderate; – adequate; – – limited; – – – poor

turn, bentonite grease is incompatible with all other types of grease.

Solid Lubricants

Solid lubricants are used especially whenever fluid and consistent lubricants cannot provide the lubricating action required. This is frequently the case under the following operating conditions: low sliding speeds, oscillating motions, high specific loads, high or low operating temperatures, extremely low ambient pressures (vacuum), and aggressive ambient atmospheres. Solid lubricants are also used to improve particular properties of fluid and consistent lubricants, i. e., as additives, for example, to minimize friction and wear and to guarantee antiseizure performance. Solid lubricants in the form of powders, pastes or lubricating varnishes contribute directly to the build up of the lubricating film on the one hand or improve the lubricating properties in oils, greases or bearing materials on the other hand.

Substances with a layer lattice structure (graphite, the sulfides MoS_2 and WS_2), selenides (WSe_2), organic substances [polytetrafluoroethylene (PTFE), amides, imides], soft nonmetals (lead sulfide, iron sulfide, lead oxide, and silver iodide), soft nonferrous metals (gold, silver, lead, copper, and indium), and reaction layers on the surface (oxide, sulfide, nitride, and phosphate layers) are used as solid lubricants. Graphite needs water to adhere and to reduce shear strength (low friction) and hence is unsuitable for use in a dry atmosphere or vacuum. Molybdenum disulfide (MoS_2) adheres well to all metal surfaces with the exception of aluminum and titanium. It is a highly suitable solid lubricant for temperatures up to 350 °C but costs more than graphite. Polytetrafluoroethylene (PTFE or Teflon) exhibits a low friction factor at low speeds and high loads and is suitable for temperatures from −250 °C to +250 °C.

Their high proportion of solid lubricants (graphite, molybdenum disulfide or PTFE) distinguishes lubricating varnishes from decorative industrial varnishes. They can be used as a dry film at temperatures between −180 °C and +450 °C. Lubricating varnishes with oil-resistant binders can also be used in oily systems and are suitable, for example, for bypassing the critical break-in phase without damage or for shortening the break-in time.

Table 5.15 Compatibility of types of lubricating grease [5.16]

Grease type	Na	Li	Ca	Ca complex	Ba complex	Al complex	Bentonite
Na grease		–	–	–	+	–	–
Li grease	–		+	+	+	–	–
Ca grease	–	+		–	+	–	–
Ca complex	–	+	+		+	–	–
Ba complex	+	+	+	+		+	–
Al complex	–	–	–	–	+		–
Bentonite	–	–	–	–	–	–	

+ compatible; – incompatible

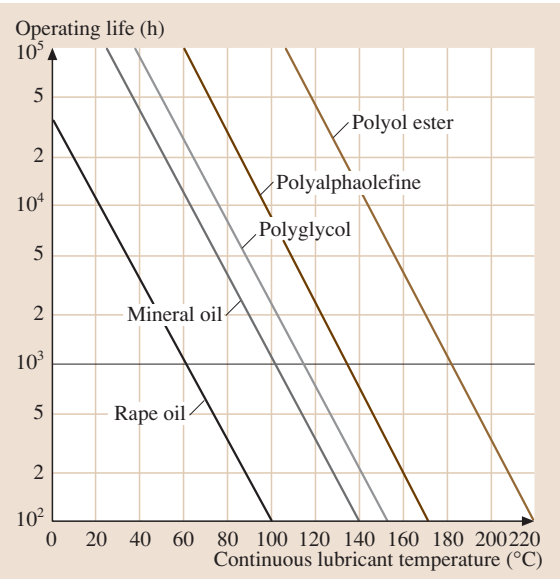


Fig. 5.18 Influence of the lubricant continuous temperature on the approximate useful life of mineral-oil-based and synthetic lubricants (after [5.17])

Properties of Lubricants

The properties of fluid and consistent lubricants are specified by data predominantly determined with standardized test procedures. It should be noted that not all results from laboratory tests are meaningful for lubrication applications. Along with viscosity, other properties of lubricants are density, specific heat, thermal conductivity, pour point, flashpoint, fire point, foaming behavior, compatibility with sealing materials, and consistency in the case of lubricating grease. Furthermore, aging resistance is vital since it characterizes the abatement of the lubricity and thus the useful life of lubricating oils and determines the oil change interval (Fig. 5.18).

Viscosity. One of a lubricant's most important rheological properties is its viscosity. A fluid's dynamic (or absolute) viscosity η is a measure of the amount it resists a relative motion. The dynamic viscosity η is defined as the shear force F required to realize the relative movement between two planes, acting in the direction of the lubricant flow between the two planes and related to the friction area A and the velocity gradient $\mathrm{d}v/\mathrm{d}y$ between the planes (Fig. 5.19).

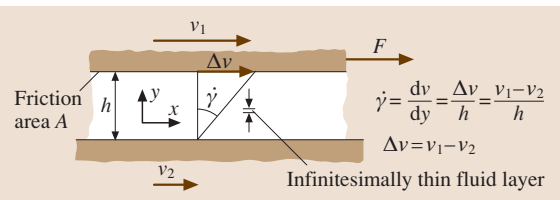


Fig. 5.19 Gap flow and velocity gradient for a Newtonian fluid

Since the shear force per friction area corresponds to the shear stress τ and the velocity gradient to the local shear strain rate $\dot{\gamma}$ (also called shear rate), the following relationship applies

$$\eta = \frac{F/A}{\mathrm{d}v/\mathrm{d}y} = \frac{\tau}{\dot{\gamma}}\,, \tag{5.8}$$

with $\dot{\gamma} = \mathrm{d}v/\mathrm{d}y = \Delta v/h$, where Δv stands for the relative velocity between the two friction bodies and h is the lubrication gap height. The absolute value of the shear force F equals that of the frictional force F_f.

Fluids that can be characterized with (5.8) at constant temperatures and pressures are also called *Newtonian fluids*. Many common fluids, especially those with relatively simple molecular structures, fall into this group (e.g., undoped mineral oils, synthetic fluids, vegetable oils, water, and gases). The unit for dynamic viscosity η is $\mathrm{N\,s/m^2} = \mathrm{Pa\,s}$. Engineering generally uses mPa s however, which corresponds to cP (centipoise), which was common earlier.

The viscosity is measured with commercially available viscosimeters, which are standardized as rotation, capillary, falling ball, and falling rod viscosimeters. Rotation and falling ball viscosimeters determine the dynamic viscosity η, while the capillary viscosimeter used most determines the ratio of the dynamic viscosity η and the density ρ. This ratio is known as the kinematic viscosity ν, so that the following applies

$$\nu = \frac{\eta}{\rho}\,. \tag{5.9}$$

The kinematic viscosity ν is a mathematical value, i. e., it is not a material property. Its units are $\mathrm{m^2/s}$. Usually however $\mathrm{mm^2/s}$ is used, corresponding to cSt (centistokes), which was used earlier. The kinematic viscosity has become generally established in industry and commerce to designate the viscosity of lubricants.

The viscosity of substances that do not behave like a Newtonian liquid is dependent on the temperature, pressure, shear rate, and mean molecular weight. In addition, shear strains are dependent not only on the instantaneous shear rate but also on the past shear history (the lubricant's *memory properties*).

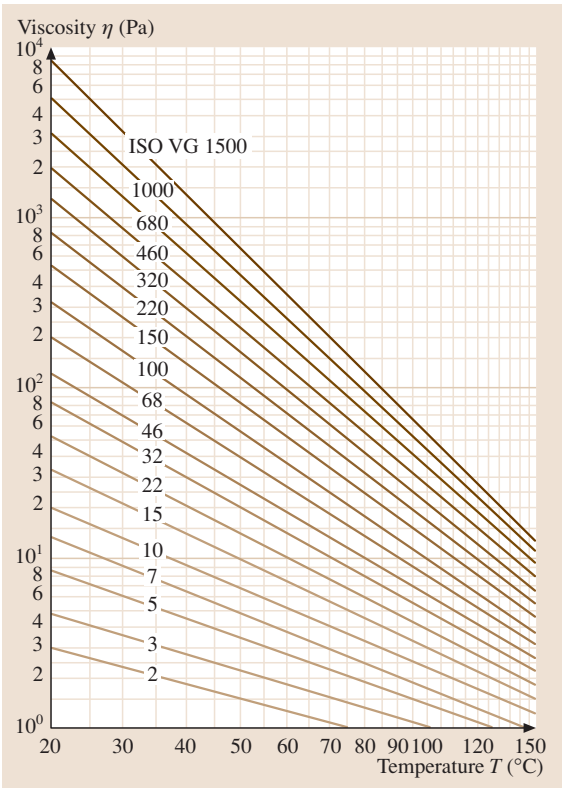


Fig. 5.20 Dependence of the dynamic viscosity η on the temperature T at density $\rho = 900\,\mathrm{kg/m^3}$ in accordance with ISO

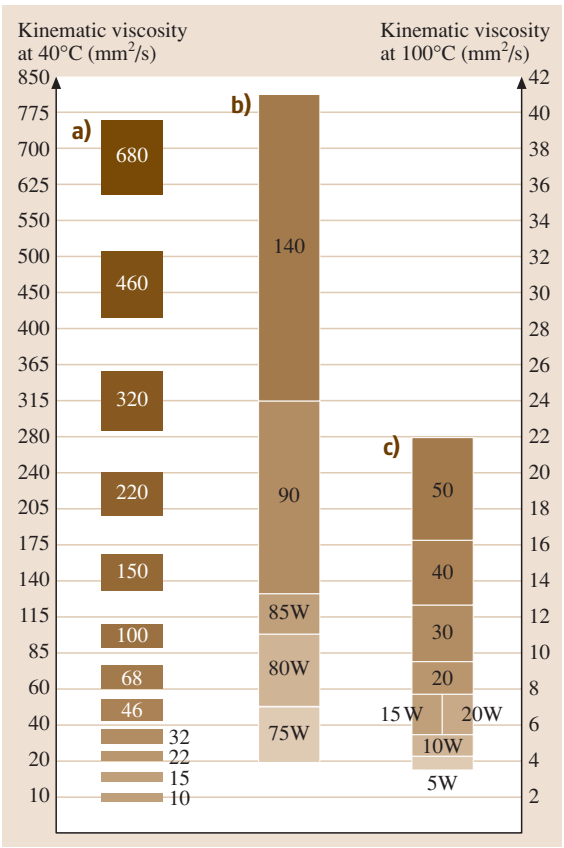


Fig. 5.21a–c Viscosity classifications in accordance with ISO and SAE [5.20] **(a)** Industrial lubricating oils (ISO VG), **(b)** SAE gear oils, **(c)** SAE crankcase oils

Dependence of Viscosity on Temperature and Pressure. The viscosity of lubricating oils greatly depends on the operating temperature. As the temperature increases, the viscosity of the lubricating oil decreases considerably.

To design lubricated tribological contacts it is important to know the viscosity at the operating temperature since this decisively influences the lubrication film thickness between the surfaces to be separated. The lubricant's viscosity–temperature behavior (η–T behavior) is determined metrologically with viscosimeters and is frequently specified by simple power and exponential equations. The following equation from Vogel has proven to be valuable for lubricating oils used in the field

$$\eta(T) = A \exp\left(\frac{B}{C+T}\right) . \qquad (5.10)$$

In this equation, η stands for the dynamic viscosity in Pa s, A, B, and C are values specific to lubricants that have to be determined for every lubricant, and T is the operating temperature in °C. Figure 5.20 presents the viscosity–temperature behavior for industrial fluid lubricants subdivided into 18 viscosity classes (ISO VG) in ISO 3448. The selection of a logarithmic scale on the ordinate axis and a hyperbolic scale on the abscissa gives the viscosity curves a straight gradient. This makes it possible to determine the η–T behavior with only two measurements.

The ISO viscosity index (VI) is also frequently used in practice to specify viscosity–temperature behavior. The viscosity index is a measure of the slope of the *straight lines* of viscosity–temperature compared with a reference lubricant. The higher the VI, the more favorable the η–T behavior. A high VI is characterized by a relatively mild change of viscosity as temperature changes and a low VI by a relatively intense change. Oils that hardly tend to thicken at low temperatures and

Table 5.16 Viscosity–pressure coefficient α and examples of increasing viscosity for different lubricants [5.21]

Type of oil	$\alpha_{25\,°C} \times 10^8$ (m²/N)	$\frac{\eta_{2000\,bar}}{\eta_0}$ at 25 °C	$\frac{\eta_{2000\,bar}}{\eta_0}$ at 80 °C
Paraffin basic mineral oils	1.5–2.4	15–100	10–30
Naphthene basic mineral oils	2.5–3.5	150–800	40–70
Aromatic solvent extracts	4.0–8.0	1000–200 000	100–1000
Polyolefines	1.3–2.0	10–50	8–20
Ester oils (diester, dendritic)	1.5–2.0	20–50	12–20
Polyehteroils (aliph.)	1.1–1.7	9–30	7–13
Silicon oils (aliph. subst.)	1.2–1.4	9–16	7–9
Silicon oils (arom. subst.)	2.0–2.7	300	–
Chlorinated paraffin (depending on level of halogenation)	0.7–5.0	5–20 000	–

that do not become low viscosity too fast at high temperatures, i. e., oils with a high VI, should be preferred when machine parts to be lubricated have to operate over a wide range of temperatures. Common paraffin basic oils exhibit a VI of 90–100, synthetic lubricants have a VI of approximately 200 and above.

Figure 5.21 provides a comparison of the viscosity classification according to SAE and ISO for different areas of application.

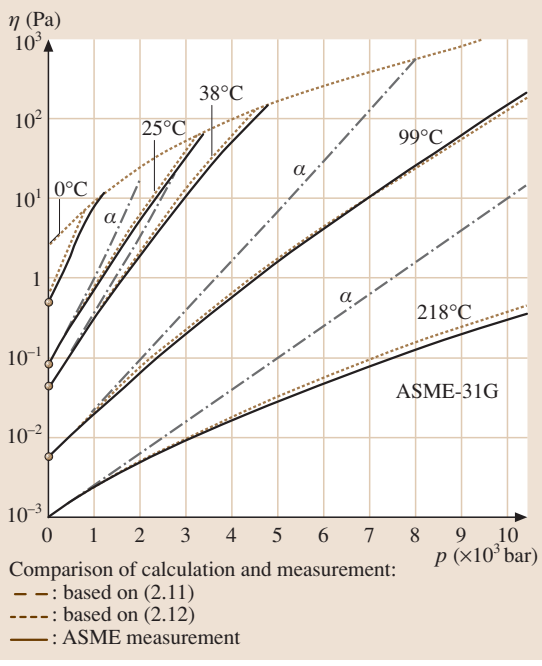


Fig. 5.22 Viscosity of the lubricating oil ASME-31G as a function of pressure and temperature [5.20] (ASME-31G is roughly equivalent to a lubricating oil ISO VG 46)

The viscosity of lubricating oils increases as the pressure rises. The viscosity's pressure dependence only really becomes noticeable at high pressure though. The influence of the pressure decreases as the temperature increases. The viscosity of mineral oil increases more strongly as the pressure increases as the viscosity–temperature curve becomes steeper. The viscosity–pressure behavior can be approximated by the following equation

$$\eta(p) = \eta_0(T) \exp\left[\alpha\left(p - p_u\right)\right] . \qquad (5.11)$$

In the equation, $\eta_0(T)$ is the viscosity at 1 bar and the corresponding operating temperature, α is the

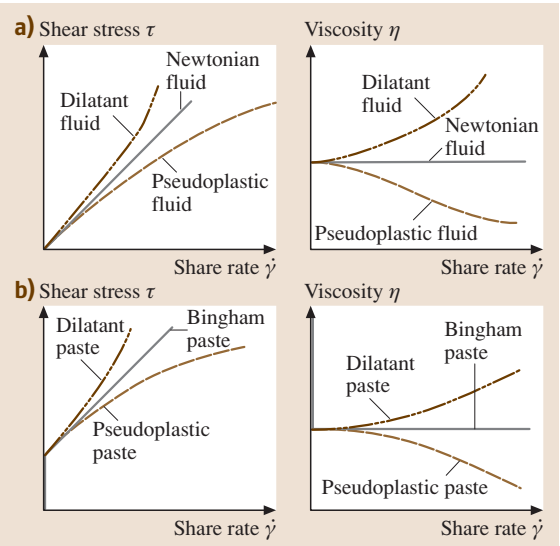


Fig. 5.23a,b Typical flow curves of different lubricants. **(a)** Newtonian, dilatant and pseudoplastic fluid; **(b)** Bingham paste, dilatant and pseudoplastic paste

Table 5.17 NLGI consistency classes and applications of lubricating greases (after [5.2]) (NLGI – National Lubricating Grease Institute)

NLGI-class	Penetration 0.1 mm	Consistency	Sliding bearings	Ball bearings	Centralized lubricating systems	Gears	Water pumps	Block greases
000	445–475	Almost fluid			+	+		
00	400–430	Semifluid			+	+		
0	355–385	Extra soft			+	+		
1	310–340	Very soft			+	+		
2	265–295	Soft	+	+				
3	220–250	Medium	+	+				
4	175–205	Relatively firm		+			+	
5	130–160	Firm					+	
6	85–115	Very firm						+
+ Primary fields of application								

viscosity–pressure coefficient, and p_u is the ambient pressure; α has a new characteristic value for every lubricant and is chiefly influenced by the composition (paraffin–naphthene–hydrocarbons and aromatics content) as well as the base oil's physical properties but less by chemical additives (Table 5.16).

Reference [5.1] provides an expression that simultaneously reproduces the dependence of the dynamic viscosity η on the state variables pressure p and temperature T

$$\eta(T, p) = A \exp\left[\frac{B}{C+T}\left(\frac{p-p_u}{2000}+1\right)^{\left(D+E\frac{B}{C+T}\right)}\right]. \tag{5.12}$$

The dependence of the dynamic viscosity on the temperature is represented by the coefficients A, B, and C (Vogel equation) and the dependence on the pressure is described by the coefficients D and E. Tests are employed to determine the coefficients A–E. Figure 5.22 presents the viscosity of a lubricating oil as a function of pressure and temperature.

Dependence of Viscosity on Shear Rate. When the rheological properties are independent of time, the flow properties of viscous lubricants can be easily described. Then the shear stress τ in the lubricant is a simple function of the local shear rate $\dot{\gamma}$, i. e., $\tau = f(\dot{\gamma})$. If this function is linear so that the shear stress is proportional to the shear rate, then a Newtonian fluid exists and the proportionality coefficient is the dynamic viscosity, which also remains constant when shear rates vary (Fig. 5.23a). Pure mineral oils generally exhibit Newtonian properties up to relatively high shear rates of 10^5–$10^6\ s^{-1}$. At higher shear rates, which occur relatively often in tribotechnical contacts such as toothed gears, ball bearing, cam-follower pairs, etc., the viscosity's constancy frequently disappears and the viscosity decreases as the shear rate increases. The lubricant begins to behave like a non-Newtonian fluid, i. e., the viscosity now depends on the shear rate.

Pseudoplastic behavior, also known as shear thinning, is characterized by a decrease of viscosity as the shear rate increases (Fig. 5.23a). Dilatant fluids manifest the opposite of pseudoplastic behavior, i. e., thickening of the lubricant as the shear rate increases (Fig. 5.23a). Dilatant fluids are normally suspensions with a high solid content.

The flow properties of greases can be compared with those of a Bingham substance. In order to generate a flow, a threshold shear stress must first be overcome (Fig. 5.23b). This means that grease behaves like a solid at first. Once the threshold shear stress τ_0 is exceeded, the lubricating grease then flows, for example, with constant viscosity like a Newtonian fluid or even pseudoplastically or dilatantly.

Consistency of Lubricating Greases. The behavior of a lubricating grease is frequently described by its consistency (plasticity). Penetration according to ASTM D-217 and ASTM D-1403 is used as a characteristic. To determine the penetration, the penetration depth of a standard cone with predetermined dimensions into the surface of a lubricating grease is measured in a penetrometer after a penetration time of 5 s at a temperature of 25 °C (in units of 1/10 mm). A difference

is made between unworked and worked penetration. Unworked penetration is measured in the unused lubricating grease, whereas worked penetration in measured in already sheared grease that has been worked under standardized conditions in a standard lubricating grease mixer. The higher the worked penetration, the softer the grease. Table 5.17 shows the relationship between penetration and consistency class.

References

5.1 Gesellschaft für Tribologie e.V.: *GfT Arbeitsblatt 7: Tribologie – Verschleiß, Reibung, Definitionen, Begriffe, Prüfung* (GfT, Moers 2002), in German

5.2 H. Czichos, K.-H. Habig: *Tribologie-Handbuch; Reibung und Verschleiß*, 2nd edn. (Vieweg, Wiesbaden 2003), in German

5.3 S. Engel: *Reibungs- und Ermüdungsverhalten des Rad-Schiene-Systems mit und ohne Schmierung*, Dissertation (Universität Magdeburg 2002), in German

5.4 A. Gervé, H. Oechsner, B. Kehrwald, M. Kopnarski: Tribomutation von Werkstoffoberflächen im Motorenbau am Beispiel des Zylinderzwickels, FVV-Heft **R**, 497 (1998), in German

5.5 J.A. Greenwood, J.B.P. Williamson: The contact of nominally flat surfaces, Proc. R. Soc. A **295**, 300–319 (1966)

5.6 B.J. Hamrock: *Fundamentals of Fluid Film Lubrication* (McGraw-Hill, New York 1994)

5.7 J.W. Kragelski: *Reibung und Verschleiß* (VEB Technik, Berlin 1971), in German

5.8 K.-H. Habig: Tribologie. In: *Dubbel – Taschenbuch für den Maschinenbau*, 21st edn., ed. by K.-H. Grote, J. Feldhusen (Springer, Berlin, Heidelberg 2004), in German

5.9 G. Fleischer, H. Gröger, H. Thum: *Verschleiß und Zuverlässigkeit* (Verlag Technik, Berlin 1980), in German

5.10 K. Wächter: *Konstruktionslehre für Maschineningenieure* (Verlag Technik, Berlin 1989), in German

5.11 H. Thum: *Verschleißteile* (Verlag Technik, Berlin 1992), in German

5.12 D. Bartel: Berechnung von Festkörper- und Mischreibung bei Metallpaarungen, Dissertation, Universität Magdeburg (2001), in German

5.13 O.R. Lang, W. Steinhilper: *Gleitlager* (Springer, Berlin, Heidelberg 1978), in German

5.14 P. Deyber: Möglichkeiten zur Einschränkung von Schwingungsverschleiß,. In: *Reibung und Verschleiß von Werkstoffen, Bauteilen und Konstruktionen*, ed. by H. Czichos (Expert-Verlag, Grafenau 1982), p. 149, in German

5.15 G. Poll: *Wälzlager: Dubbel – Taschenbuch für den Maschinenbau*, 21st edn. (Springer, Berlin, Heidelberg 2004), in German

5.16 U.J. Möller, J. Nassar: *Schmierstoffe im Betrieb*, 2nd edn. (Springer, Berlin, Heidelberg 2002), in German

5.17 G. Niemann, H. Winter, B.-R. Höhn: *Maschinenelemente Band 1; Konstruktion und Berechnung von Verbindungen, Lagern, Wellen*, 3rd edn. (Springer, Berlin, Heidelberg 2001), in German

5.18 W.J. Bartz: Additive – Einführung in die Problematik Kontakt und Studium. In: *Additive für Schmierstoffe*, Vol. 433, ed. by W.J. Bartz (Expert, Renningen-Malmsheim 1994), in German

5.19 G.W. Stachowiak, A.W. Batchelor: *Engineering Tribology*, 2nd edn. (Butterworth-Heinemann, Boston 2001)

5.20 Gesellschaft für Tribologie e.V.: *GfT-Arbeitsblatt 5: Zahnradschmierung* (GfT, Moers 2002), in German

5.21 D. Klamann: *Schmierstoffe und verwandte Produkte. Herstellung-Eigenschaften-Anwendung* (VCH, Weinheim 1982), in German

6. Design of Machine Elements

Oleg P. Lelikov

A machine generally consists of a motor, a drive, and an actuating element. The mechanical power driving a machine constitutes the rotary motion energy of a motor shaft. Electric motors, internal-combustion motors, or turbines are the most common types of motors. The mechanical power transmission from the motor to the actuating element is accomplished by various driving gears. These include gearings, worm gearings, belt drives, chain drives, and friction gears. Some examples of actuating elements are car steering wheels, work spindles, and screw propellers of ships. This chapter covers the advanced design of machine elements, in particular all common types of gearings and the needed machine components. The in-depth description including stress and strength analysis, materials tables and assembly recommendations allows for a comprehensive and detailed calculation and design of these most important drives. Shafts and axles, shaft-hub assemblies and bearings are included with design guidelines and machining options. Single machine elements, such as specific information about bolts and bolted joints, springs, couplings and clutches, friction drives and also sliding bearings are dealt with only where needed for the benefit of a more general view. The chapter provides the practicing engineer with a clear understanding of the theory and applications behind the fundamental concepts of machine elements.

6.1 Mechanical Drives

In general, three components (Fig. 6.1) can be specified in a machine: a motor, a drive, and an actuating element. The mechanical power driving a machine is the rotary motion energy of a motor shaft. Electric motors, internal-combustion motors, or turbines are the most common types of motors. The mechanical power transmission from the motor to the actuating element is carried out through different driving gears (hereinafter referred to as *gears*), including: gearings, worm gearings, belt drives, chain drives, and friction gears. Some examples of actuating elements are car steering wheels, work spindles, and the screw propellers of ships. A machine without a gear would be optimal; as examples we can consider electrical spindles, electric motors with grinding wheels on the shafts, which do not have transmission gears. This lack of a transmission gear is due to the coincidence of the required rotational work frequency of the grinding wheel and that of the motor shaft. However, in practice this coincidence occurs very rarely [6.1–26].

Gears are applied in order to obtain the required power and kinematic parameters of the actuating element.

Depending on the operating mode, the following mechanical drives can be distinguished:

- Gears working through meshing (gearings, worm gears, chain drives)
- Gears working through friction (friction gears, belt drives)

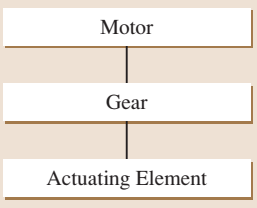


Fig. 6.1 Structural arrangement of a machine

When transmitting power, mechanical gears can simultaneously carry out one or more of the following functions:

1. *Decrease (or increase) the rotational frequency* from the motor shaft to the actuating element shaft (Fig. 6.2).

   The main parameters of the drive shaft (index 1) and the driven shaft (index 2) are: power P_1, P_2 (kW), torque T_1, T_2 (N m), and rotational frequency n_1, n_2 (min^{-1}).

   The torque T (N m) of every shaft can be calculated using the power P (kW) and rotational frequency n (min^{-1}) as

$$T = 9550P/n \; .$$

   Obviously, a decrease in the rotational frequency leads to a torque increase, and an increase in the rotational frequency leads to a torque decrease.

   An important characteristic of gears is the gear ratio u, determined as the ratio of the rotational frequencies of the n_1 drive shaft and the n_2 driven shaft or (without taking into account sliding in contact) as a diameter ratio of the driven (d_2) and drive (d_1) gear elements (Fig. 6.2)

$$u = n_1/n_2 = d_2/d_1 \; .$$

   If $u > 1$ the rotational frequency of the driven shaft is less than that of the drive shaft by a factor of

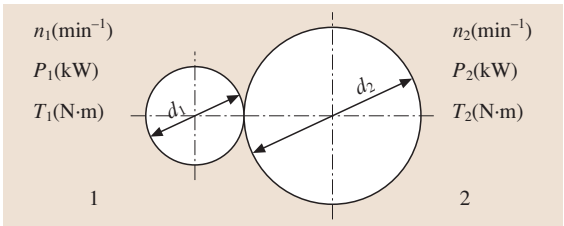


Fig. 6.2 Main parameters on the drive shaft (1) and driven shaft (2) of the gear

the gear ratio:

$$n_2 = n_1/u\ .$$

The decrease of rotational frequency is called reduction, and the enclosed drive decreasing the rotational frequency is called a reduction gear. Units subjected to increasing rotational frequency are called accelerators or multiplying gears. Hereinafter we will consider only reduction gears, which are more common in applications.

In most cases, in practice the rotational frequency of the motor shaft considerably exceeds the rotational frequency of the actuating element shaft; for example, the shaft rotational frequency of the internal-combustion engine of a car is 5000 min^{-1}, whereas the wheel rotational frequency due to the rate of car movement is 100 km/h (1000 min^{-1}).

The ratio of powers and torques: The power P_2 of the driven shaft is less than the power P_1 of the drive shaft due to friction losses in the gear, which are quantified by the efficiency factor η

$$P_2 = P_1\eta\ .$$

The torque of the driven shaft increases by a factor of the gear ratio (in accordance with the decrease of its rotational frequency)

$$T_2 = T_1 u\eta\ .$$

2. *Change of the power flow direction*: A back-axle gearing serves as an example in this case. The motor shaft axis of rotation in many cars forms an angle of 90° with the rotation axis of the wheels. For the transmission of mechanical power between shafts with intersecting axes, bevel gearings are used (Fig. 6.3).
3. *Speed regulation of the driven shaft*: The value of the torque changes with the variation of rotational frequency; a higher torque corresponds to a lower frequency. The necessity for a high torque, for a car for example, appears when starting motion or when driving on a slope. This is also the case, for example, when a lathe is removing thick metal chips. For speed regulation of the driven shaft, gearboxes and variable-speed gears can be used.

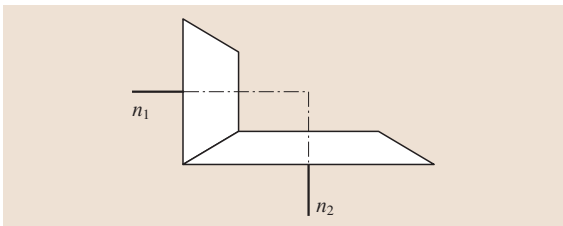


Fig. 6.3 Bevel gearing with intersecting shaft axes

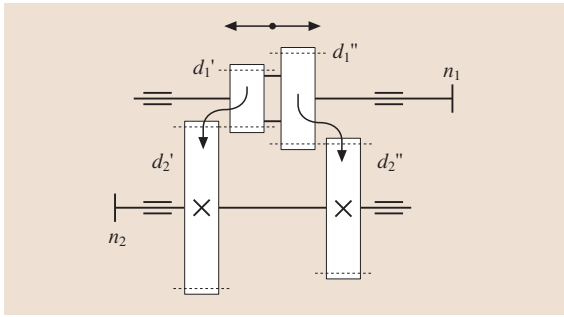


Fig. 6.4 Double gear cluster

Gearboxes ensure staging of the rotational frequency of the driven shaft, depending on the number of stages and the covered stage. For the double-stage gearbox (Fig. 6.4) we have

$$u_1 = n_1/n_2 = d_2'/d_1' \quad \text{and}$$
$$u_2 = n_1/n_2 = d_2''/d_1''\ .$$

Variable-speed gears provide an infinitely variable change of the rotational frequency of the driven shaft over a certain range. Variable-speed friction drive units with a power of up to 55 kW are mainly used in mechanical drives.

In the front variable-speed gear (Fig. 6.5) the change of the rotational speed of the driven shaft is achieved by moving a small roller (1) along the shaft, i. e., through a distance R_i towards the driven shaft axis. The gear ratio u_i falls in the range from

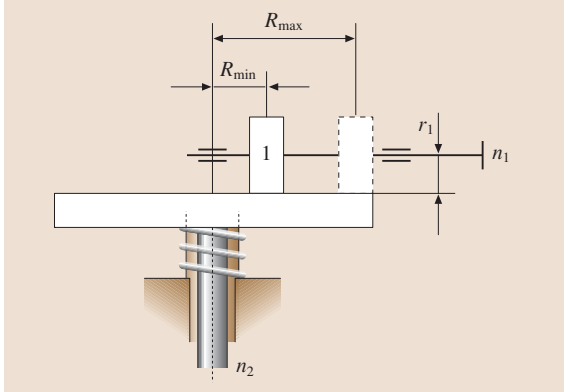


Fig. 6.5 Front variable-speed drive

$u_{min} = R_{min}/r_1$ to $u_{max} = R_{max}/r_1$. Hence the adjustment range is

$$D = u_{max}/u_{min} .$$

For the front variable-speed gear $D \approx 2.5$.

4. *Transformation of one kind of movement to another* (rotational movement to translational movement, uniform to discontinuous motion, etc.).
5. *Motion reversal* (in stroke and return motion).
6. *Motor power distribution* between a few actuating elements of the machine.

6.1.1 Contact Stresses

Contact stresses occur due to the interaction of bodies whose contact area dimensions are small in comparison with the dimensions of the contacting bodies themselves, for example, the contact of two circular steel cylinders along the *common generating line* (an analog of a toothing friction gear, and frictionless roller bearings); see Fig. 6.6. Due to the influence of the external force, when gears and rolling-contact bearings move, the contact occurs over small areas (an initial contact along a line or at a point); as a result, high stresses arise in the surface layer, and the material around this area suffers from volumetric stress.

In the field of the mechanics of contact interaction, the classical work of the German physicist Henry Hertz *About the Contact of Solid Elastic Bodies* (1882) was one of the first publications to appear.

For the initial fingertip action the direct force F_n is distributed throughout the contact area as a pressure diagram representing a semi-ellipsoid (in this particular case, a hemisphere). The pressure p_0 has a maximum value in the center of the contact area.

According to Hertz's theory the maximum *direct stress* acts in the center of the contact area and is equal to the maximum pressure p_0 in absolute value. These stresses are called *Hertz's stresses* and are identified as: $\sigma_H = p_0$. The maximum contact stress is denoted by the index "H", which stands for Hertz.

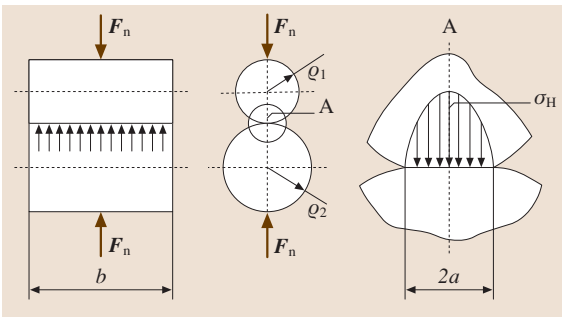


Fig. 6.6 Stress distribution in the contact of circular cylinders with parallel axes

The contact of cylinders without a load occurs at parallel axes along the generating line. Under the influence of a compressive force F_n, as a result of the cylindrical elastic strain, the initial contact along the line becomes a contact over an elongated rectangular area (a very narrow zone) whose width $2a$ is significantly smaller than its length b.

For an initial linear contact, the direct force F_n is distributed throughout the contact area as a pressure diagram representing an elliptic semicylinder. The pressures in the profile throughout the contact area change according to the elliptic law and reach a maximum value σ_H in the zone of peak deformations along the line of the compressive force action (remote element A). This particular feature of direct contact stress action is due to the fact that the stresses do not spread deeply into the bodies of the cylinders, being concentrated in a thin surface layer.

The contact area dimensions and the arising direct stresses σ_H depend on the loading F_n, the elastic material characteristics (Poisson's ratio and the coefficients of elasticity), and the form of the contact bodies.

In addition to the direct stress σ_H there are also tangential stresses τ in the contact area. The maximum tangential stress $\tau_{max} = 0.3\sigma_{H\,max}$ acts at a point located on the line of the compressive force action F_n at a distance of $0.78a$ from the contact surface.

The values of the contact stresses considerably exceed those of other stresses (tension, bending), as well as the material mechanical characteristics in the uniaxial stress state: σ_y, σ_t. Thus, in frictionless bearings $\sigma_H = 4600\,N/mm^2$, while for steel grade 100 Cr 6 (used in Europe) the yield strength is $\sigma_y = 1700\,N/mm^2$, and the ultimate strength is $\sigma_t = 1900\,N/mm^2$. The absence of destruction in the presence of such high stresses can be explained by the fact that in the zone of action the material is under overall volumetric compression.

The maximum value σ_H is used as the main criterion for the contact strength

$$\sigma_H \le [\sigma]_H ,$$

where $[\sigma]_H$ is an allowable contact stress determined from experiment or operating experience with similar conditions in the contact zone.

For the calculation of the maximum contact stress on the contact area, Hertz's formula is applied, obtained

from the contact solution of elasticity theory as

$$\sigma_H = \sqrt{\frac{1}{\pi\left[\left(1-\nu_1^2\right)/E_1 + \left(1-\nu_2^2\right)/E_2\right]}} \times \sqrt{\frac{F_n}{b}\sum\left(1/\rho_i\right)}\,, \qquad (6.1)$$

where b is the contact line length (the length of the cylinders), υ_1 and υ_2 are the Poisson's ratios of the contacting bodies materials, E_1 and E_2 are their coefficients of material elasticity, and ρ_1 and ρ_2 are the radii of curvature of the contact areas.

For the contact of two convex surfaces (Fig. 6.7a)

$$\sum(1/\rho_i) = 1/\rho_1 + 1/\rho_2\,,$$

whereas for the contact of a convex and a concave surface (Fig. 6.7b)

$$\sum(1/\rho_i) = 1/\rho_1 - 1/\rho_2\,,$$

so that in general this can be written

$$\sum(1/\rho_i) = 1/\rho_1 \pm 1/\rho_2\,.$$

Hertz's formula is derived under the following conditions:

- The materials of the contacting bodies are uniform and isotope.
- The compressive forces are directed along the straight line joining the centers of curvature of the body surface at the point of initial contact and are such that only elastic strains take place in the contact zone.
- Frictional forces in a contact are absent.
- Body surfaces are absolutely smooth and have an ideal form.
- There is no lubricant on the contact surfaces.
- The cylinder length is infinite.

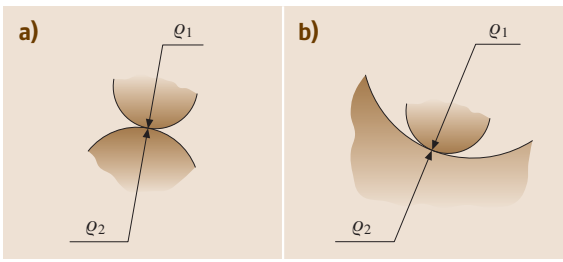


Fig. 6.7a,b Surface contacts: **(a)** convex and **(b)** convex and concave

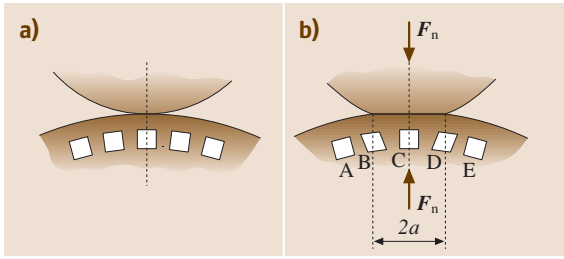


Fig. 6.8a,b Deformations of microvolumes of material due to the rolling of cylinders: **(a)** without load and **(b)** under load

The contact line length of real products is finite, frictional forces act on the contact area, and the surfaces are lubricated. The applicability of the stated Hertz formula is allowed by the fact that the allowable stresses $[\sigma]_H$ are found experimentally for conditions similar to the operating conditions of the design product.

Deformations of the material microvolume in the contact zone with cylinder rolling are sketched in Fig. 6.8a (without load) and b (with load). The material of each body contacted by free rolling is subject to repeated loading and unloading as it passes through the strain area (Fig. 6.8). Here the assigned material microvolume experiences a cycle of reversible slip and compression A–B–C–D–E. Nevertheless, the material behaves as an ideal elastic body in macrovolumes under a relatively small load.

6.1.2 Nature and Causes of Failure Under the Influence of Contact Stresses

Rumpling of Contact Surfaces

The rumpling of the contact surfaces occurs due to impacts as well through the application of vibrational loads, or under the action of considerable loads, when plastic strains occur in addition to elastic ones.

Fatigue Flaking

Every point on the surface during the rotation of the cylinders experiences the action of the contact stresses σ_H (point A in Fig. 6.9a), and the surface itself experiences strain cycling. The fatigue crack (2) arising from repeated microplastic slips appears near the surface (1) of the cylinder (Fig. 6.10a), on the site of stress concentration through the presence of micro-irregularities or nonmetallic inclusions, which are always present in steel.

Within the distorted layer the crack spreads obliquely to the surface and then along the border of the

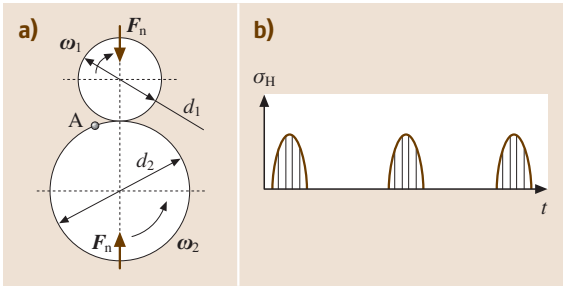


Fig. 6.9a,b Repeated loading at the point A **(a)** with contact stresses σ_H **(b)**

distorted layer. Fatigue crack propagation into deeper layers is connected with the wedging action of the lubricant.

The lubricant (3) before the contact area under the action of the high pressure developing in the hydrodynamic layer is discharged into the crack opened by the frictional forces (2) (Fig. 6.10b). Within the contact area under the loading the crack is closed and increased pressure of the lubricant is produced (Fig. 6.10c), which assists in crack propagation until the abruption of the metallic particle (4) from the surface (Fig. 6.10d). This first occurs for shallow cavities (the size of a fraction of a millimeter) and then through chipping of their edges, after which the cavities join together to form large cavities with a typical flaw dimension of 2–5 mm. Flaking violates the formation conditions for a continuous oil film (i.e., oil squeezes into the cavities), which results in surface wear and tear.

Due to the small thickness of the hardened case and considerable contact stresses, cracks can arise in the depth of the material under the hardened case or on the border of the hardened case as a result of bond cleavage on the borders of metal grains. The imbalance of intracrystalline bonds results in flaking of the hardened case.

Microplastic slips, the range of which depends on the hardness of the material, underlie material accumulation of fatigue damage. The contact strength increases with increasing hardness.

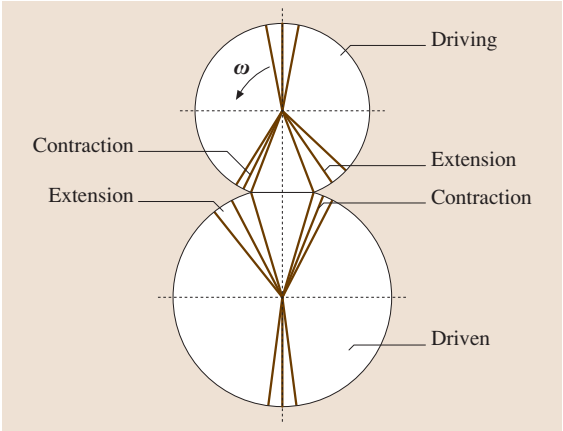


Fig. 6.11 Hoop strains on the drive and driven cylinders

Wear

Frictional forces in the contact area produce compressive strain in the circle line on the surface of the *drive cylinder* before the contact area, and they produce tensile strain after it. On the driven cylinder the opposite is true: before the contact area there is tensile strain, and after it there is compressive strain. These strains are schematically shown in Fig. 6.11 by the angular separations of the radii. When passing through the contact area a relative point displacement of the drive and driven cylinders is observed, i. e., *relative slip*, which is the cause of wear.

Seizing

In the absence of lubricants or in the case of breakage of an oil layer under strong loading, relative slip results in a considerable local temperature increase and molecular adhesion (microwelding) with subsequent breaking and picking up of torn-out parts of the material onto the corresponding mated surface.

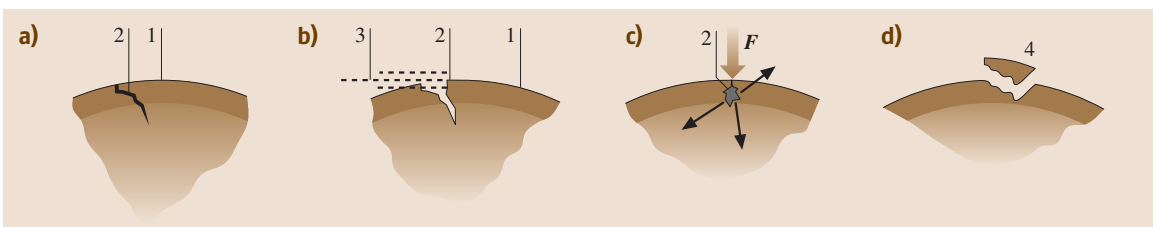


Fig. 6.10a–d Fatigue flaking. **(a)** Initiation of fatigue crack, **(b)** delivery of lubricant into the open with frictional forces in the crack, **(c)** closing of the crack under loading, and **(d)** removal of the metallic particle

6.2 Gearings

6.2.1 Basics

In gearings, movement is transmitted using the toothing of gear wheel pairs. The smaller gear wheel is called a *pinion* and a larger one is called a *wheel*. The term *gear wheel* refers to both the pinion and the wheel.

Advantages of gearings are:

1. Relatively small dimensions and mass of gear wheels, with high load-carrying capacity and safety
2. High efficiency factor (97–98%)
3. Applicability over a wide range (circumferential forces from 0.01 N in instrumental devices to 1000 kN in rolling mill drives)
4. Applicability over a wide velocity range (circumferential velocity from about 0 m/s in systems for telescope displacement to 250 m/s in the drive of helicopter rotors)
5. Comparatively small loads on the shafts and bearings
6. Persistence of the average value of the gear ratio
7. Easy maintenance

Disadvantages of gearings are:

1. The necessity for high accuracy in manufacture and mounting
2. Noise when operating gearings with high rotational frequencies, which results from varying tooth rigidity and imprecision of step and teeth profiles

Wheel teeth must be obtained through cutting or rolling.

Gearings are applied in a wide range of fields and working conditions, for example, in watches and measuring instruments, transmissions of cars, tractors, other vehicles, and road-building machines, lifting and steering crane devices, machine gearboxes, drives of rolling mills, carriers, etc.

Gearings are divided according to the form of the pitch surface into a number of types: cylindrical gearings (with external or internal toothing), and bevel gearings.

Cylindrical Gearings with External Toothing

In these gearings the toothing is equivalent to rolling without slip of cylinders with diameters of d_{w1} and d_{w2}, which are called the initial diameters. In gearings with external toothing the initial surfaces of the gear wheels are located one outside the other. The pinion in the reducing gear is a drive element and all its characteristics are indicated by the index 1, for example, rotational frequency n_1 (min^{-1}) and teeth number z_1. The characteristics of the driven element of the wheels are referenced by the index 2: n_2 and z_2 (Fig. 6.12).

The intersection lines between the teeth side faces and any circumferential cylindrical surface coaxial with the initial surface are called *teeth lines*. If teeth lines are parallel to the gear wheel axis this is called a *spur* (Fig. 6.12a). If these lines are helical with constant pitch the gear wheel is described as *helical* (Fig. 6.12b). With

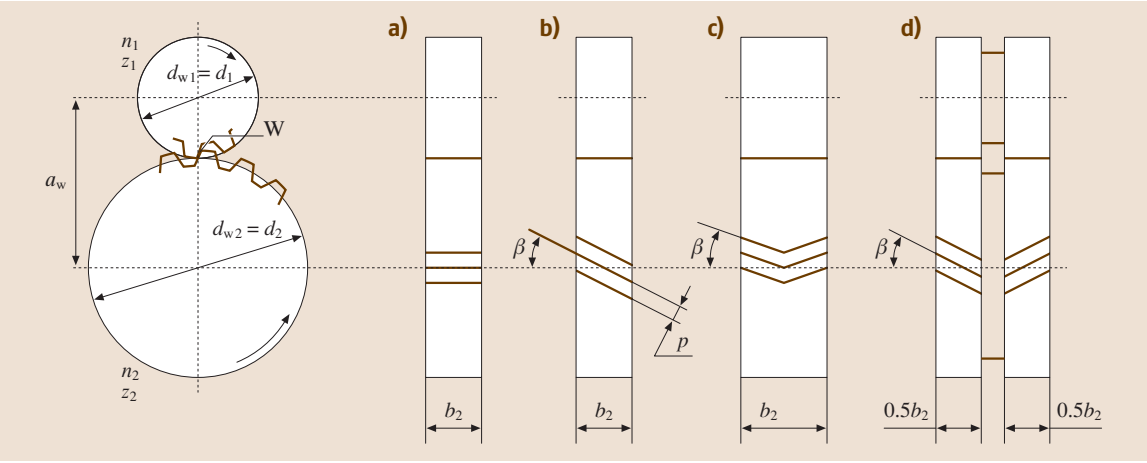


Fig. 6.12a–d Cylindrical gearing with external toothing. **(a)** Straight-toothed, **(b)** helical, **(c)** and **(d)** herring-bone, respectively, with and without a groove as an outlet of the tooth-cutting tool

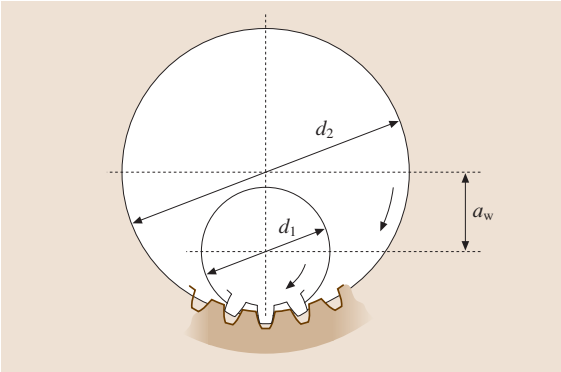


Fig. 6.13 Cylindrical gearing with internal toothing

increasing tooth angle tilt β the load-carrying capacity of the gearing increases, but an axial force arises, which influences bearings and shafts (usually $\beta = 8-18°$).

Another type of helical gear wheel is the herringbone gear, either without a flute (Fig. 6.12c) or with a flute for run-out of tooth-cutting instruments (Fig. 6.12d). In such a gear, as a consequence of the teeth being tilted in opposite directions on the semi-chevrons, axial forces are mutually balanced on the wheel and do not load the bearing; normally $\beta = 25-40°$.

The tangent point W of the initial circumferences of the pinion d_{w1} and the wheel d_{w2} is called the *pitch point*.

The pitch surface (pitch cylinder) is a cylinder where the gear wheel spacing is equal to the spacing of the original profile, i. e., to the spacing of the generating rack. For simplicity of presentation we will consider gearings without displacement so that the initial d_w and pitch d diameters of their gear wheels coincide: $d_1 = d_{w1}$ and $d_2 = d_{w2}$. However, in designating the axle base the index "w" is retained, as in a_w.

The distance between analogous profile points of adjacent teeth on the pitch diameter measured in the section normal to the teeth line is called a *normal spacing* p. The ratio p/π is called a *module* and is designated by m. The module is the main characteristic of the teeth dimensions. It is measured in millimeters and is chosen from a standard series 2, 2.5, 3, 4, etc.

The main features of a gearing can be written using the characteristics of the gear wheels as follows:

- The gear ratio, taking into account that $d = mz$,

$$u = n_1/n_2 = d_2/d_1 = z_2/z_1$$

- The axle base $a_w = 0.5(d_2 + d_1)$

The width b_2 of the gear wheel is normally less than the width of the pinion. In calculations the ratio $\psi_{ba} = b_2/a_w$, known as the *width coefficient*, is used. The values ψ_{ba} of are standardized to 0.1, 0.125, 0.16, 0.2, 0.25, 0.315, 0.4, 0.5, 0.63, and 0.8. For gearboxes, narrow wheels ($\psi_{ba} = 0.1–0.2$) are applied to decrease the dimensions in the direction of the shaft axes, and wide reduction gear wheels (with $\psi_{ba} = 0.315–0.63$) are used.

Cylindrical Gears with Internal Toothing

In gears with internal toothing (Fig. 6.13) the initial surfaces of the gear wheels are located one inside the other. In this case, the axle base is $a_w = 0.5(d_2 - d_1)$.

In general,

$$a_w = 0.5(d_2 \pm d_1)\,,$$

where the plus sign refers to gears with external toothing and the minus sign refers to gears with internal toothing.

Compared with gearings with external toothing, gearings with internal toothing have smaller dimensions

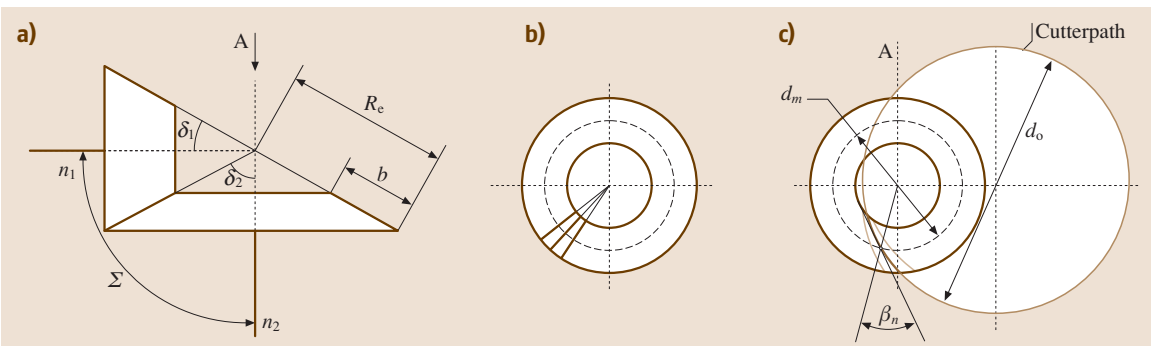


Fig. 6.14 (a) Bevel gearing with intersecting shaft axes, (b) with straight teeth, and (c) with circular teeth

and mass, and operate more smoothly due to the higher contact ratio and the fact that teeth contacting at convex and concave surfaces have a larger equivalent curvature radius. Moreover, they have a lower slip velocity.

Bevel gearings transmit mechanical power between shafts with intersecting axes. Normally, $\Sigma = \delta_1 + \delta_2 = 90^\circ$ (Fig. 6.14a). The toothing of the bevel wheels can be considered as a rolling of the pitch circular cones of the pinion and the wheel. The main characteristics of bevel gearings are the angles of the pitch cones, δ_1 and δ_2, and the external cone distance R_e. Intersection lines of the teeth side faces with the pitch cone surface are called *teeth lines*. Depending on the form of the tooth line there are gearings with *straight teeth* (Fig. 6.14b), where teeth lines go through the vertex of the pitch cone, and *circular teeth* (Fig. 6.14c), which are circular arcs d_0.

Bevel wheels with circular teeth are characterized by the tooth line tilt in the middle section according to the width of the gear ring. The *tilt angle* β_n is the acute angle between the tangent to the tooth line and the generation of the pitch cone (Fig. 6.14c).

Another version of bevel gearings is the *hypoid* gearing, where the rotation axes of the gear wheels do not intersect but cross.

6.2.2 Accuracy of Gearings

The working capacity of gearings depends considerably on the production accuracy of the gear wheels. Production errors are unavoidable due to: deviation in pitch, profile, tooth direction; radial run-out of the gear ring; deviation from parallelism and misalignment of the gear wheel axes; center distance variation; etc. These errors result in increased noise, loss of rotational accuracy of the driven wheel, failure of precision and smooth toothing, torsional vibration, dynamic increase and decrease of distribution evenness along the contact line acting in the load toothing, and other detrimental effects.

Standards regulate the accuracy of gear wheels as well as cylindrical and bevel gearings. Twelve degrees of accuracy are specified and are designated in decreasing accuracy order by the numbers from 1 to 12. Most often, degrees 6, 7, and 8 are applied, where degree 6 corresponds to high-accuracy speed gears, degree 7 corresponds to gears with a normal grade of accuracy that operate with high speed and moderate load, or with moderate speed and large load, and degree 8 corresponds to low-accuracy gears. Gears rated for manufacture according to the sixth degree of accuracy can have a mass of the gear set that is 30% less than that required with the eighth degree of accuracy. For each degree of accuracy there are three standards of tolerances, which are detailed below.

The standard for kinematic accuracy regulates the difference between the actual and nominal rotation angles of the driven gear wheel. The indices of kinematic accuracy influence external gearing dynamics and the positional accuracy of the output shaft with respect to the input shaft. Because of the risk of torsional and resonance oscillations, and noise, these are important in the pitch circuits of machines, control systems, and high-speed power trains.

The standard for smooth operation regulates rotary speed fluctuations per wheel revolution, which cause high-frequency variable, dynamic loads, and noise.

The standard for teeth contact regulates the teeth adjacency in the mounted gearing and the degree of load distribution in contact lines, and determines the efficiency of power trains.

The gearing *side clearance* is also regulated. This is the distance between the teeth side faces, which determines the free rotation of one of the gear wheels by a fixed double gear wheel. Side clearance is required to avoid teeth seizing in the gearing as a result of their expansion at the working temperature, as well as to provide a location for lubricant and for the provision of free-wheel rotation. Side clearance is provided in conjunction with tolerances of teeth thickness and the axle base. The clearance dimension is specified by a coupling type of gear wheel in the gearing: H = 0 clearance, E = small, D and C = reduced, B = standard, A = increased. Mostly coupling types B and C are applied. For reverse gears it is recommended to use couplings with reduced clearances. An example of the accuracy designation of a cylindrical gearing with grade 7 according to the standards of kinematic accuracy, grade 6 according to the standards of drive operation smoothness, grade 6 in accordance with the standards of teeth contact, and with coupling type C is 7-6-6-C.

6.2.3 Gear Wheel Materials

The choice of the gear wheel material is made to provide contact strength and teeth bending resistance for the functioning gearing under its operating conditions. Steel is the most commonly used material in power trains. In some cases cast iron and plastic are also used. The important criteria for the selection of materials are the mass and dimensions of the gearing.

The materials used for gear wheel production in Russia are discussed below. The correspondence between Russian and foreign materials is provided in Appendices 6.A and 6.B.

Steel: Gearings with steel gear wheels have the lowest mass and dimensions. Moreover, the mass and dimensions decrease with greater hardness of the teeth effective area, which in turn depends on the steel grade and the heat treatment applied.

Heat *refining* treatment is a combination of quenching and high-temperature tempering; it provides the most favorable combination of hardness, viscosity, and plasticity.

Heat refining treatment is carried out before teeth cutting. Materials for the wheels are carbon steel grades C36, C35, C46, C45 (EN), 50Г, and alloy steel grades 37Cr4 (DIN), 5145 (ASTM), 40NiCr6 (DIN), etc. The hardness of the tooth core and the tooth effective area are equal for improved wheels, 235–302 HB. Wheel teeth made from refined steel have good running in ability and are not subject to fracture failure, although they have restricted load-carrying capacity. They are applied in lightly and medium loaded gearings.

High hardness ($H > 350$ HB) of the surface layer with viscous core preservation is achieved using thermal or chemicothermal surface hardening of previously refined gear wheels. This includes surface hardening, cementation, nitrocementing with tempering, and nitriding.

Surface hardening of teeth with high-frequency current heating is appropriate for gear wheels with module values > 2 mm. For low modules a small tooth is annealed through, which results in warpage and embrittlement of the tooth. Steel grades C46, C45 (EN), 37Cr4 (DIN), 40NiCr6 (DIN), and 34CrMo4KD (DIN) are applied for quenching with high-frequency current heating; their surface hardness is 45–53 HRC. For $H > 350$ HB material hardness is measured according to the C-Rockwell scale. The tooth core hardness corresponds to the heat refining *treatment*.

Cementation (surface diffusion carburizing) with subsequent quenching along with high surface hardness also provides a high bending strength for the teeth. For gear wheels of medium size, the carburized case constitutes 15% of the tooth thickness (but not more than 1.5–2 mm). Only the surface layer saturated with hydrocarbon is annealed. Steel grades 5120 (ASTM), 14NiCr10 (5732) (DIN), and 20MnCr5G (DIN) (hardness of the tooth surface 56–63 HRC) are used for cementation.

Nitrocementing (nitrocarburizing) of the teeth surface layers in a gaseous medium with subsequent quenching provides high contact and bending strength, wear, and sliding strength. Steel grades 5120 (ASTM), 20CrMo5 (DIN), and 30MnCrTi (DIN) are applied. The nitrocarburizing layer thickness is 0.1–1.2 mm. Warpage (tooth distortion) is insignificant, and subsequent grinding is not required. The hardness of the tooth surface is 58–64 HRC

Nitriding (surface diffusion nitrogen saturation) provides particularly high hardness of the teeth surface layers. It is characterized by insignificant warpage and enables the production of teeth of high accuracy without development operations. Nitrided wheels are not used under impact loads (because of the risk of cracking the hardened case). Steel grades 41CrAlMo7 and 40NiCrMo4KD (EN) (hardness 58–65 HRC) are applied for nitrided wheels.

Strengthening heat treatment is carried out before nitriding, i. e., quenching with subsequent high-temperature tempering. The teeth are not ground after nitriding and nitrocementing because of the minimum warpage. This is why these kinds of chemicothermal hardening can be successfully used for wheels with internal teeth and in cases when teeth grinding is difficult to carry out. Nitriding is not used as often as cementation and nitrocementing due to the long process involved (several tens of hours) and the resulting thin layer (0.2–0.8 mm).

Wheel teeth with hardness $H > 45$ HRC are cut before heat treatment. *Teeth finishing* (grinding, etc.) is carried out *after the heat treatment*, as required. Gearings with hard ($H > 45$ HRC) work surfaces of the teeth run in badly.

Throughout surface heat or chemicothermal treatment of the teeth, previous heat treatment (refining) defines the mechanical characteristics of the tooth core.

The load-carrying capacity of gearings corresponding to the contact strength is higher when the surface teeth hardness is higher. Thus it is advisable to use surface thermal or chemicothermal hardening. These kinds of hardening allow one to increase the load-carrying capacity of the gearing several fold in comparison with refined steels; for example, the allowable contact stresses $[\sigma]_H$ of cemented gear wheels are twice as high as the values of $[\sigma]_H$ of heat-refined wheels, which allows one to decrease their mass by four times. However when defining the hardness of the teeth work surfaces, it must be borne in mind that *higher hardness corresponds*

to a *more difficult manufacturing technique for the gear wheels* and moderate gearing dimensions (which can result in difficulties with the structural unit).

Core-mold casting is used in the manufacture of large gear wheels ($d > 600$ mm). The steel grades used are GS52 to GS55 (DIN). Cast wheels are subjected to *normalization* [heating to 750–950 °C, soaking, and subsequent air cooling (austenization)].

Cast iron is used in manufacturing gear wheels of slow speed, large dimensions, and open gearings. The cast-iron grades used are GG20 to GG35 (DIN). The teeth of cast-iron wheels grind well and can operate with poor lubrication. They have low bending strength; therefore the dimensions of cast-iron wheels are considerably larger than those of steel wheels.

Plastics are applied in high-speed low-load gearings for wheels working together with steel or cast-iron wheels (because of the low thermal conductivity of plastic and the risk of jamming).

Plastic wheels are produced to be narrower than mating metallic wheels in order to avoid increased wear of the edges of the mating wheels. Plastic gear wheels are remarkable for their silence and running smoothness. However, they cannot be used in high-loaded gearings. The most common plastics used are textolite, caprolan, polyacetal resin, and phenylone.

6.2.4 The Nature and Causes of Gearing Failures

As the gearing operation passes through the meshing zone the teeth are subjected to repeated loading. In addition the direct force on the surface and frictional forces act on the contacting teeth surfaces. *For every tooth, stresses change in time according to a zero-to-tension stress cycle, which is the cause of fatigue damage*, which may result in flaking of the work surfaces or teeth breakage. Slipping and frictional forces during meshing produce wear and teeth jamming.

Fatigue flaking of the work teeth surfaces, which is *the main fracture mode of teeth for most closed well-lubricated gearings*, is a consequence of the cycling of the contact stresses. The fracture starts near the pitch line (1) (Fig. 6.15a), where the highest load (the zone of one-pair contact) acts. This is also the location of the highest frictional force (near the pitch line the slip velocity is minimum), which assists in the production and propagation of the microcracks and grooves (2) on the surface of the teeth. *In open gearings* (without lubrication) flaking is not observed; wear of the teeth surface surpasses fatigue crack propagation.

Bearing failure of the teeth work surface occurs under the action of considerable loads or impacts on load application.

Teeth breakdown, in comparison with work surface damage, is uncommon. However, it is the most dangerous type of failure because it results in total operational failure. Teeth breakdown belongs to the category of sudden and complete failure and is a consequence of the zero-to-tension stress cycle action of bending or overload. Fatigue breaking is related to crack propagation (3) (Fig. 6.15b) at the tooth root on the side where lower strains arise because of bending. Straight short teeth are completely broken off along the profile at the tooth root. Following fatigue failure a *concave* surface A is left on the wheel body after tooth breakage. Following damage as a consequence of overload a *convex* surface B remains.

Teeth wear is *the main type of teeth fracture observed in open gearings, gearings with a solid-lubrication coating, and gearings with a very small coating thickness* (up to 3 μm). Wear is determined by the slip ratio of the contact teeth surfaces. Frictional forces of the drive teeth are directed away from the initial circle, whereas forces of the driven teeth are directed towards the initial circle. When it becomes worn the tooth becomes thinner, its dedendum is loosened, and meshing gaps increase, which results in a loss of kinematic accuracy and, in the case of considerable wear, leads to teeth breakdown. Teeth fracture is preceded by increased noise during the operation of the gearing.

Teeth seizure is a molecular adhesion (bonding process) of the mated teeth surfaces as a consequence of

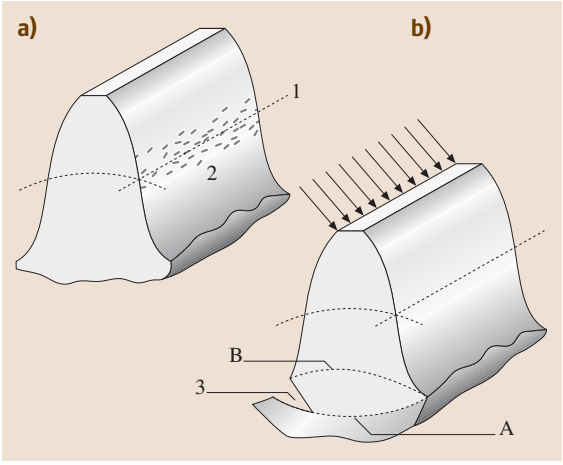


Fig. 6.15a,b Types of fatigue failure of teeth. **(a)** Flaking of the effective area, **(b)** fracture at the tooth root

damage to the lubricating film and an increase in local temperature, which is caused by the relative slip in the contact zone. Scales due to molecular adhesion tear the work surfaces of the mated teeth, furrowing them in the slip direction. Teeth seizing can be avoided through increase in hardness, reduction of the roughness of the surface of the work teeth, teeth modification, and matching of extreme-pressure oils.

6.2.5 Choice of Permissible Contact Stresses Under Constant Loading Conditions

During the operation of gearing meshing the teeth are subjected in turn to the action of a zero-to-tension stress cycle. If the cycle operation factors are constant in time, the loading conditions are called *regular*. The cycle of variable in time loading conditions is called irregular.

The loading law is described by a *sequence diagram*, which consists of a load curve (torque T, force F) as a function of operation time (or number N of loading cycles). A sequence diagram under constant loading conditions is represented in Fig. 6.16, where T_1 is the torque and N_k is the gearing lifetime in number of stress-change cycles.

The choice of permissible stresses is based on the stress-cycle diagram. *Stress-cycle diagrams* obtained experimentally on specimen prototypes of the *gear wheels* define in the coordinates σ (the highest cycle stress) and N (the number of stress changing cycles, which the specimen had before the fracture) (Fig. 6.17a). Experiments show that these diagrams have two typical areas: a left area that is sloping and a right area that is horizontal (Fig. 6.17a) or that has a slight tilt towards the cycle axis (Fig. 6.17b). A logarithmic scale is often used on the abscissa axis. In this case the tilted area of the stress-cycle diagram is substituted by a straight line (Fig. 6.17b). The limit number $N_{\lim}$ of cycles is called the *abscissa of inflection* of the stress-cycle diagram. $N_{\lim}$ is also called the base number of loading cycles. The stress-cycle diagram of the tilted area is described by the power function

$$\sigma_i^q N_i = C \,,$$

where C corresponds to the experimental conditions (material hardness, specimen dimensions, etc.).

The fatigue limit point σ_{RN} is defined by a given value N_k of the cyclic service time in the stress-cycle diagram (Fig. 6.17a), and the limit value of the cycle number N_1 before fracture is determined by the given stress level σ_1 (Fig. 6.17b). If $N_k \geq N_{\lim}$, the stress $\sigma_{\lim}$ is a fatigue point with a zero-to-compression stress cycle. The limit number $N_{\lim}$ of stress cycles corresponds to the fatigue point $\sigma_{\lim}$.

The diagrams are curved for different stress types (contact or bending), for different materials, and heat treatment types; they differ in values $\sigma_{\lim}$, $N_{\lim}$, exponent q, number C.

The stress-cycle diagram for the contact stresses is shown in Fig. 6.18, where the logarithmical coordinates have two tilted areas with exponents $q = 6$ (left) and 20 (right), respectively. The number $N_{H\,lim}$ of cycles corresponding to the change of the stress-cycle diagram is defined by the average hardness of the teeth surface according to

$$N_{H\,lim} = 30\,HB_m^{2.4} \leq 12 \times 10^7 \,.$$

The fatigue contact point $\sigma_{H\,lim}$ is calculated from empirical formulas depending on the material and type of heat treatment for the gear wheel and the average hardness H_m of the teeth surface. Thus, for the heat refining treatment: $\sigma_{H\,lim} = 2\,HB_m + 70\,N/mm^2$.

Contact stresses with a number of changing stress cycles of N_i are calculated in accordance with the equation for the stress-cycle diagram

$$\sigma_{Hi}^q N_i = \sigma_{H\,lim}^q N_{H\,lim} \,.$$

Hence

$$\sigma_{Hi} = \sigma_{H\,lim} \sqrt[q]{N_{H\,lim}/N_i} \,.$$

Then contact stresses by the specified lifetime N_k

$$\sigma_H = \sigma_{H\,lim} \sqrt[q]{N_{H\,lim}/N_k} = \sigma_{H\,lim} Z_N \,,$$

where

$$Z_N = \sqrt[q]{N_{H\,lim}/N_k} \,.$$

The specified lifetime N_k by the rotational frequency n in min^{-1} and the lifetime action period L_h in h

$$N_k = 60 n n_t L_h \,,$$

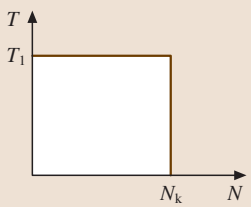


Fig. 6.16 Sequence diagram of constant loading conditions

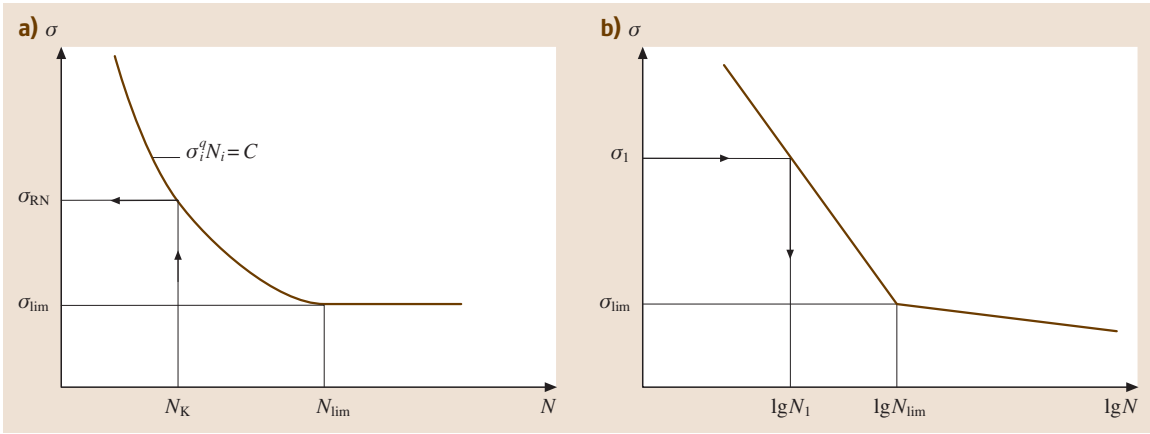


Fig. 6.17a,b Stress-cycle diagram in different coordinates: **(a)** common and **(b)** logarithmic in the abscissa axis

where n_t is the number of tooth matings of the wheel per revolution (Fig. 6.19).

Allowable stresses for the pinion ($[\sigma]_{H1}$) and for the wheel ($[\sigma]_{H2}$) are defined according to the general formula (but with substitution of the appropriate parameters for the pinion and the wheel, respectively), taking the influence of the lifetime on the contact strength, the roughness of the mated teeth surfaces, and the circumferential velocity into account

$$[\sigma]_H = \sigma_{H\,lim} Z_N Z_R Z_V / S_H \,.$$

The service life ratio Z_N takes into account the *influence of lifetime*.

For $N_k \le N_{H\,lim}$ (the left-hand area of the stress-cycle diagram)

$$Z_N = \sqrt[6]{N_{H\,lim}/N_k}\,, \qquad (6.2)$$

on the condition that $Z_N \le Z_{N\,max}$, where $Z_{N\,max} = 2.6$ for materials with homogenous structure (normalized, refined, volume-quenched), and $Z_{N\,max} = 1.8$ for surface-hardened materials (quenching, cementation, nitriding).

Inequality limits allowable stresses according to the conditions preventing of plastic strain or the brittle fracture of the surface layer.

For $N_k > N_{H\,lim}$ (the right-hand area of the stress-cycle diagram)

$$Z_N = \sqrt[20]{N_{H\,lim}/N_k}\,,$$

on the condition that $Z_N \ge 0.8$.

The ratio Z_R, taking into account the influence of the initial roughness of the mated teeth surfaces, is accepted for the gear wheel of the pair with the roughest surface, depending on the roughness parameter *Ra*

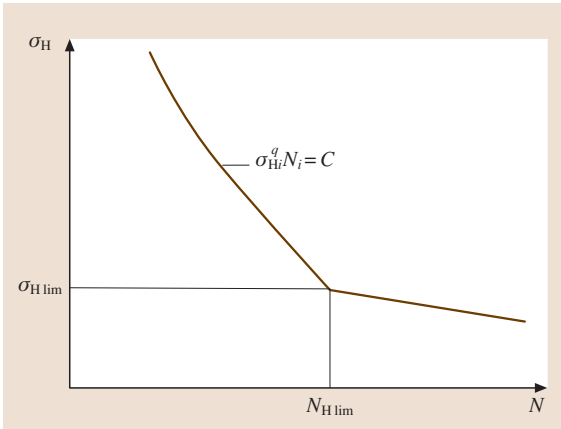


Fig. 6.18 Stress-cycle diagram for contact stresses

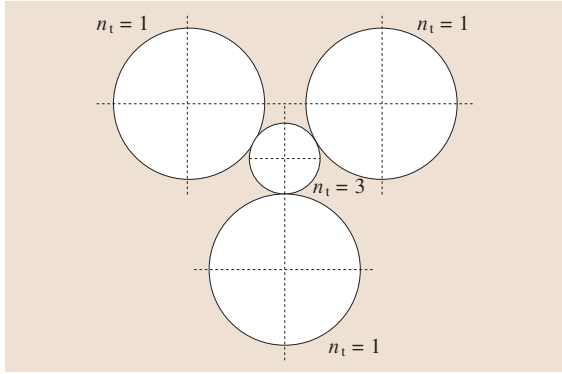


Fig. 6.19 Determination of the loading number of a tooth per revolution of the gear wheel

($Z_R = 1–0.9$). High values of Ra correspond to a polished and ground face ($Ra = 0.63–1.25\,\mu m$).

The ratio Z_V takes into account the influence of the circumferential velocity ($Z_V = 1–1.15$). Lower values correspond to fixed gearings that operate with low circumferential velocities ($v < 5\,m\,s^{-1}$). With higher values of the circumferential velocity better conditions arise for the formation of a reliable oil layer between the contact teeth areas, which enables the allowable stresses to be increased.

Minimum values of the *load factor* for gear wheels with a homogenous material structure (normalized, refined, volume-quenched) are $S_H = 1.1$; for gear wheels with surface hardening the value is $S_H = 1.2$.

The allowable stress $[\sigma]_H$ for cylindrical gearings with straight teeth is the lowest of the permissible stresses of the pinion $[\sigma]_{H1}$ and the wheel $[\sigma]_{H2}$.

For cylindrical gearings with indirect teeth as a result of the contact line location on the angle to the pitch line the allowable stresses can be increased to the value

$$[\sigma]_H = \sqrt{0.5\left([\sigma]_{H1}^2 + [\sigma]_{H2}^2\right)}\,,$$

in the case of the execution condition $[\sigma]_H \le 1.2[\sigma]_{H\,min}$, where $[\sigma]_{H\,min}$ is the lower of $[\sigma]_{H1}$ and $[\sigma]_{H2}$.

The allowable stress for bevel gearings with straight and indirect teeth is the lowest of the permissible stresses of the pinion $[\sigma]_{H1}$ and the wheel $[\sigma]_{H2}$.

6.2.6 Choice of Permissible Bending Stresses Under Constant Loading Conditions

The stress-cycle diagram for bending stress is shown in Fig. 6.20. The exponent $q = 6$ applies for normalized and refined gear wheels, while $q = 9$ applies for quenched and face-hardened teeth. The index F is attached to all the parameters concerned with the bending stress calculation performed for the tooth foot. The cycle number corresponding to the change of the stress-cycle diagram is $N_{F\,lim} = 4\times10^6$. The fatigue point $\sigma_{F\,lim}$ in the case of the zero-tension stress cycle is taken from experimental data or is calculated in agreement with an empirical formula. Thus, for heat refining treatment: $\sigma_{F\,lim} = 1.75\,HB_m$, where HB_m is the average hardness of the tooth core.

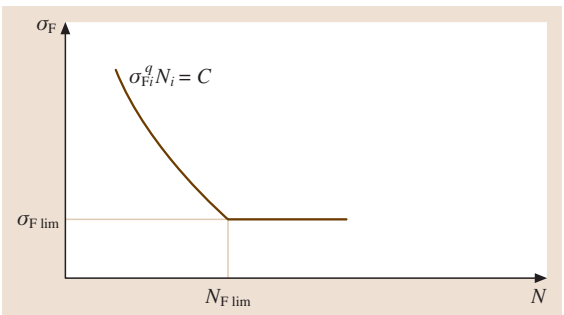


Fig. 6.20 Stress-cycle diagram for bending stresses

The bending stress by the cycle number of stress changing N_i is calculated in accordance with the equation of the stress-cycle diagram (Fig. 6.20)

$$\sigma_{Fi}^q N_i = \sigma_{F\,lim}^q N_{F\,lim}\,.$$

Hence

$$\sigma_{Fi} = \sigma_{F\,lim}\sqrt[q]{N_{F\,lim}/N_i}\,.$$

Then bending stresses with the given lifetime N_k

$$\sigma_F = \sigma_{F\,lim}\sqrt[q]{N_{F\,lim}/N_k} = \sigma_{F\,lim}Y_N\,,$$

where $Y_N = \sqrt[q]{N_{F\,lim}/N_k}$ on the condition that $Y_N \ge 1$. The given lifetime N_k is calculated in the same way as in contact stresses calculations.

In compliance with the stress-cycle diagram stresses σ_F cannot have values lower than $\sigma_{F\,lim}$, which is why $N_k > N_{F\,lim}Y_N = 1$.

Allowable stresses $[\sigma]_{F1}$ for the pinion and $[\sigma]_{F2}$ for the wheel are assigned according to the general relation (but with substitution of the appropriate parameters for the pinion and the wheel) taking into account the influence of the lifetime on the bending fatigue resistance, the surface roughness of the concave molding (the fillet surface between the adjacent teeth), and the reversing gear (for double-sided applications) of the load

$$[\sigma]_F = \sigma_{F\,lim}Y_N Y_R Y_A / S_F\,.$$

Application of modern hardening methods (knurling, cavity stamping or electropolishing) allows an increase in the breaking strength of approximately 25%.

The service life ratio Y_N takes into account the influence of lifetime

$$Y_N = \sqrt[q]{N_{F\,lim}/N_k}\,, \qquad (6.3)$$

with $1 \le Y_N \le Y_{N\,max}$, where $Y_{N\,max} = 4$ for normalized and refined wheels, and $Y_{N\,max} = 2.5$ for quenched and surface-hardened wheels.

For long-term running (a few years) of high-speed gearings, $N_k \ge N_{F\,lim}$, and therefore $Y_N = 1$, which corresponds to the first inequality sign in (6.3). The second

inequality sign limits the allowable stresses according to the condition for the prevention of plastic strain and tooth brittle fracture.

The ratio Y_R, which takes into account the influence of the roughness of the fillet surface between the teeth, is assigned the value $Y_R = 1$ in the case of grinding and gear milling with roughness parameter $Rz \leq 40\,\mu m$, and $Y_R = 1.05–1.2$ in the case of polishing (high values with refining and after quenching with high-frequency current).

The ratio Y_A takes into account the influence of the double-sided load application (reversing gear). In the case of one-sided load application $Y_A = 1$. In the case of reversing loading (reversed cycle) $Y_A < 1$.

The minimum value of the load factor is $S_F = 1.7$, whereas for cemented and nitrocemented gear wheels it takes the value $S_F = 1.55$.

6.2.7 Choice of Permissible Stresses Under Varying Loading Conditions

Most gearings operate under varying loading conditions in which the cycle parameters, e.g., the load value and consequently the stress value, change in time.

In Fig. 6.21a a gearing loading sequence is characterized by the moment sequence diagram, where the torques T_i that act during the functioning of the target life N_k are represented in decreasing order. With the help of the torque sequence diagram n_{ki} can be determined, a period (measured in loading cycles) of the torque operation T_i with rotational frequency n_i, and N_{ci}, a period (in loading cycles) of the torque operation that does not exceed T_i.

A sequence diagram of the moments can be represented in increasing order of the torques and by application of relative units (Fig. 6.21b) $\nu_i = T_i/T_{max}$ and $n_{ci} = N_{ci}/N_k$, where $N_{ci} = (n_{ki})$ is a cumulant (accumulated sum) of the loading cycle numbers. The abscissa n_{ci} in Fig. 6.21b corresponds to the part of the general loading cycle with relative torque that does not exceed the value of ν_i. The period of influence of moments higher than ν_i is characterized by the relative cycle number $(1 - n_{ci})$.

In operation the maximum possible torques (e.g., on starting) are short-term (single) and are not taken into consideration in calculations of fatigue resistance. In calculations of fatigue resistance the actual varying loading conditions are changed into equivalent (in terms of fatigue action) constant conditions.

Let us suppose that the component is operating under varying loading conditions, which have some

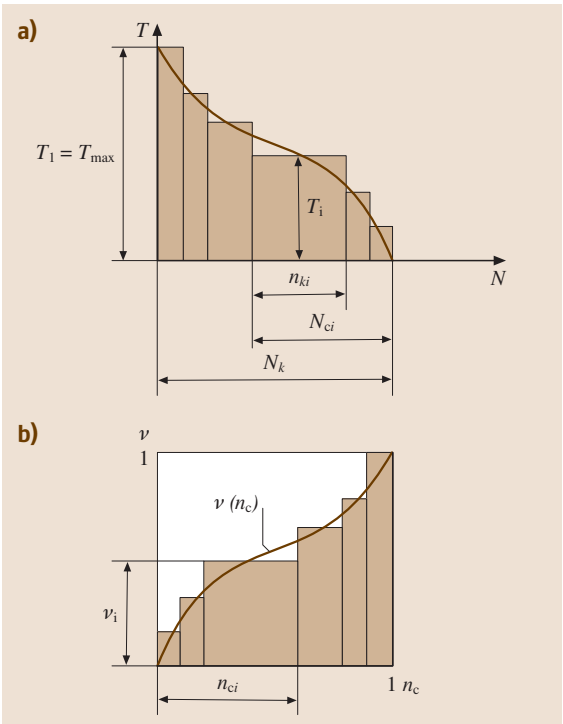


Fig. 6.21a,b Sequence diagram of the varying loading conditions in coordinates. **(a)** Absolute, **(b)** relative

phases, and that at phase i it experiences a number of fluctuation cycles n_{ki}. It is known from experience that detail fracture under periodical loading with constant parameters of the stress cycle (operating on the same phase) follows the stress-cycle diagram (Fig. 6.22) after N_i loading cycles as a result of step-by-step accumulations of damage in the material (e.g., minute cracks). Then the damage level of the detail operating on the phase i can be evaluated by using the relative life n_{ki}/N_i.

It has been determined experimentally that, by working on a few loading phases, damage continues to accumulate independently in proportion to the appropriate relative life. Thus, it can be summed up *arcwise* (the hypothesis of linear summing up of fatigue damage). In this case, fracture will take place when the sum of the relative service life is 1

$$\sum (n_{ki}/N_i) = 1\,.$$

Let us do the transformations: multiply and divide the expression with the index of summation by σ_i^q

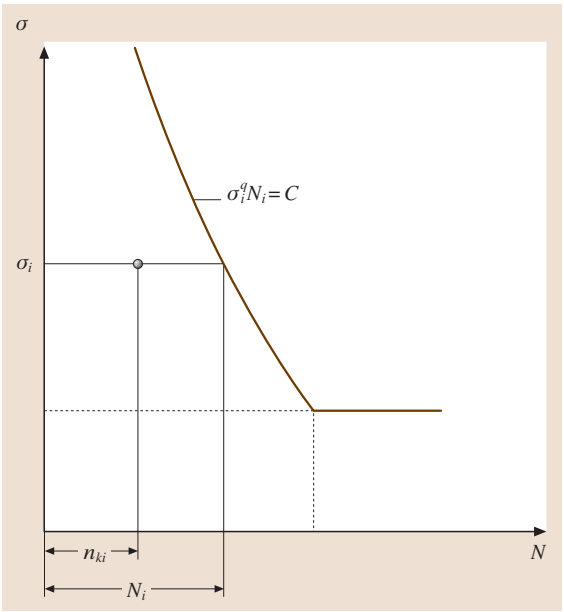


Fig. 6.22 Determination of the relative service life

stresses corresponding to the loading lever i

$$\sum \frac{\sigma_i^q n_{ki}}{\sigma_i^q N_i} = 1 \, .$$

In accordance with the equation of the stress-cycle diagram, $\sigma_i^q N_i = C$ is a constant value and can be taken out of the index of summation

$$\sum \left(\sigma_i^q n_{ki}\right) = \sigma_i^q N_i \, ,$$

and then written using parameters σ^q and N_E of the equivalent constant loading conditions

$$\sum \left(\sigma_i^q n_{ki}\right) = \sigma_i^q N_i = \sigma^q N_\mathrm{E} \, .$$

In other words, the actual varying loading conditions can be associated with the equivalent constant conditions, when the detail reaches the same degree of fatigue damage. As an equivalent the constant conditions with the nominal torque T (the highest of the long-term ones $T = T_1 = T_{\max}$ in Fig. 6.23) that generates stresses σ and with equivalent number N_E of loading cycles. From the last ratio we obtain a relation for the calculation of the equivalent number of stress change cycles

$$N_\mathrm{E} = \sum \frac{\sigma_i^q}{\sigma^q} n_{ki} \, ,$$

where $n_{ki} = 60 n_\mathrm{t} n_i L_{hi}$ is the number of stress change cycles on the loading level i for L_{hi} hours of work.

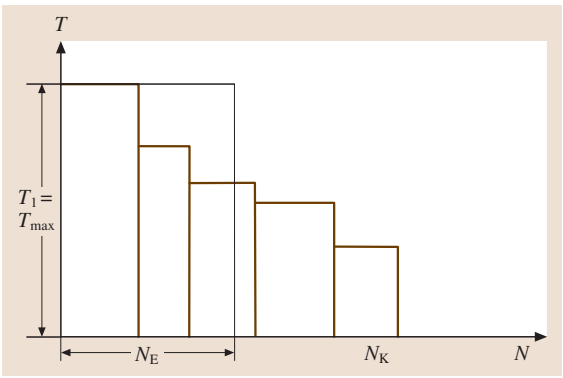


Fig. 6.23 Conformity of the equivalent constant loading conditions with real varying loading conditions

After the right-hand side of the given equation is multiplied and divided by $N_k = \sum n_{ki}$, it can be written as

$$N_\mathrm{E} = \sum \left(\frac{\sigma_i}{\sigma}\right)^q \frac{n_{ki}}{N_k} N_k = \mu N_k \, ,$$

where $\mu = \sum (\sigma_i/\sigma)^q \, n_{ki}/N_k$ is a reduction ratio.

Because contact stresses are proportional to the degree 0.5 load, and bending stresses are proportional to the degree 1 load, by substituting stresses through the torques in the expression for the reduction ratio we have

$$\mu_\mathrm{H} = \sum \left(\frac{T_i}{T_{\max}}\right)^{q/2} \frac{n_{ki}}{N_k} \quad \text{for contact stress,}$$

$$\mu_\mathrm{F} = \sum \left(\frac{T_i}{T_{\max}}\right)^{q} \frac{n_{ki}}{N_k} \quad \text{for bending stress.}$$

The equivalent numbers N_HE and N_FE of loading cycles needed to calculate the contact and bending strength are determined as

$$N_\mathrm{HE} = \mu_\mathrm{H} N_k; \quad N_\mathrm{FE} = \mu_\mathrm{F} N_k \, .$$

Calculation of the allowable stresses under varying loading rates is performed according to the formulas for the constant loading rate with substitution in agreement with (6.2) and (6.3) of the service life ratio Z_N and Y_N values of cycle numbers N_k for equivalent cycle numbers N_HE and N_FE, respectively. *Thus load variability is taken into account by the choice of the allowable stresses.*

6.2.8 Typical Loading Conditions

On the basis of statistical loading analysis of different machines it has been ascertained that the wide

variety of torque (loads) sequence diagrams can be brought together to a few *typical* ones by using the relative coordinates $\nu, (1-n_c)$ in the sequence diagram construction. Having substituted a step sequence diagram for a smooth enveloping curve (Fig. 6.21a,b) a graphical presentation of constant (zero) and five typical loading conditions that are characteristic for most modern machines (Fig. 6.24) is obtained. In this figure the varying rates are: I = *heavy* (most of the time loads are close to nominal values); II = the *average equiprobable* (the same working time with all load values); III = the *average normal* (most of the time under operation with average loads); IV = *light* (most of the time operation with loads below the average); and V = *especiallylight* (most of the time under operation with light loads).

The heavy rate (I) is typical of gearings of mining machines, the average equiprobable (II) and normal (III) rates are typical for vehicles, and light (IV) and especially light (V) rates are typical for versatile machine tools.

The typical rates (heavy, light, and especially light) can be described mathematically with integral beta distribution functions with appropriate parameters. The average equiprobable rate corresponds to a uniform distribution, and the average normal rate is of normal distribution. The equivalence ratios μ_H and μ_F are initial points of order k of the load distribution function. The order k of the initial point is determined according to the exponent q of the stress-cycle diagram equation and to the type of stress.

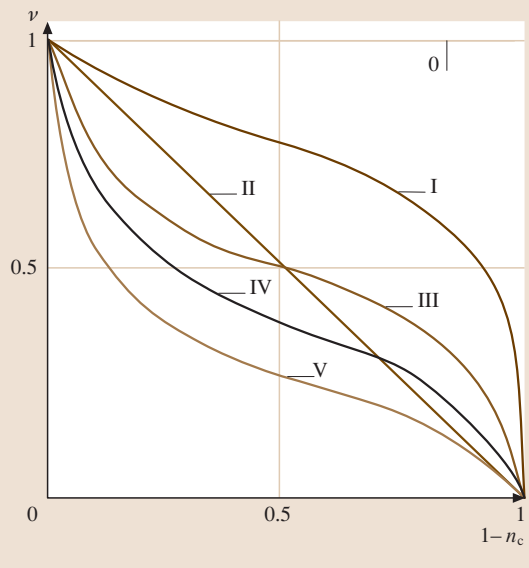


Fig. 6.24 Typical loading conditions: 0 – constant, I – heavy, II – average equiprobable, III – average normal, IV – light, and V – especially light

The values of the equivalence ratios μ_H and μ_F for the typical loading rates are calculated and given in GOST 21354-87 (in Russia) [6.27–37].

A typical rate nearest to the actual rate in the range of high loads can be taken as the calculation rate. The use of typical rates considerably simplifies these calculations.

6.2.9 Criteria for Gearing Efficiency

For well-lubricated gearings operating in the closed case the main efficiency criteria are contact strength and bending strength. *Contact strength* is the capability of the contacting teeth surfaces to provide the required safety against progressive fatigue flaking. Calculation of the prevention of fatigue failure results in the condition

$$\sigma_H \le [\sigma]_H ,$$

where σ_H is the contact stress at the pitch point, and $[\sigma]_H$ is the allowable contact stress. Calculation of the prevention of bearing failure in the case of overload results in the condition

$$\sigma_{H\,max} \le [\sigma]_{H\,max} ,$$

where $\sigma_{H\,max}$ and $[\sigma]_{H\,max}$ are, respectively, the actual and allowable contact stresses under the action of the peak load (e.g., under operation).

The bending strength is the capability of the teeth to provide the required safety against tooth fatigue fracture. Calculation for the prevention of fatigue fracture results in the condition

$$\sigma_F \le [\sigma]_F ,$$

where σ_F is the bending stress in a weak section and $[\sigma]_F$ is the allowable bending stress of a tooth.

Calculation of the prevention of overload breakage results in the condition

$$\sigma_{F\,max} \le [\sigma]_{F\,max} ,$$

where $\sigma_{F\,max}$ and $[\sigma]_{F\,max}$ are, respectively, the actual and allowable bending stresses under the action of the peak load.

The aim of gearing calculations is efficiency in accordance with all the considered criteria. In the planning calculation the gearing geometry is determined by the

given loading conditions. In the verification calculation the load-carrying capacity or compliance with the main efficiency criteria are determined by the known gearing parameters. Most often the dimensions of the closed gearing are determined through contact strength analysis, and bending analysis of the teeth is checked in order to determine the minimum possible module value.

In the case of very high teeth hardness ($H \geq 56$ HRC) gearing dimensions are determined through bending analysis of the teeth and verified by contact strength analysis.

6.2.10 Calculated Load

Due to the additional load of the gearing, additional loads act on the toothing. These are caused by loading conditions; manufacturing errors; yielding of the teeth, shafts, and bearings; and production errors of the single parts comprising a gearing unit. In calculations this is taken into consideration by multiplication of the nominal torque T or force F by the load ratio K, thereby determining the calculated load according to

$$T_c = KT \quad \text{or} \quad F_c = KF\,.$$

In contact stress analysis the load ratio is

$$K_H = K_A K_{H\beta} K_{HV} K_{H\alpha}\,.$$

With the ratio K_A the environmental dynamic load of the gearing is taken into account by the combined action of the motor and the actuating element, which is not taken into consideration in the loading sequence diagram. The value of K_A depends on the degree of loading evenness of the motor and the actuating element ($K_A \geq 1$).

Typical loading characteristics of motors are *even* (electric motors, and steam and gas turbines) and *with average unevenness* (multicylinder internal-combustion engines). The typical loading rates of the actuating element are *even* (evenly functioning belt and apron conveyors) and with *slight unevenness* (the same conveyors for piece loads). For an even loading rate of the motor and loading rate of the actuating element with slight unevenness, $K_A = 1.25$. If environmental dynamic loads are taken into account in the loading sequence diagram then $K_A = 1$.

The index of the $K_{H\beta}$ unbalance factor along the contact lines is chosen because load distribution unevenness is caused by changing the initial tilt angle β of the tooth. The ratio K_{HV} takes into account internal loading dynamics, which is caused by circular pitch errors and teeth profile faults of the pinion and the wheel. The index emphasizes the main influence on K_{HV} circumferential velocity value.

The index of the $K_{H\alpha}$ load distribution coefficient between the teeth in connection with circular pitch errors and tooth direction is caused by the fact that the load distribution between the teeth is considered in the normal plane where the angle of action a is measured.

Load Distribution Unevenness Along the Contact Lines (*Factor* $K_{H\beta}$)

Deviation from the contact line position is the result of manufacturing errors (tooth direction error) and elastic strains of the shafts and bearings. This deviation

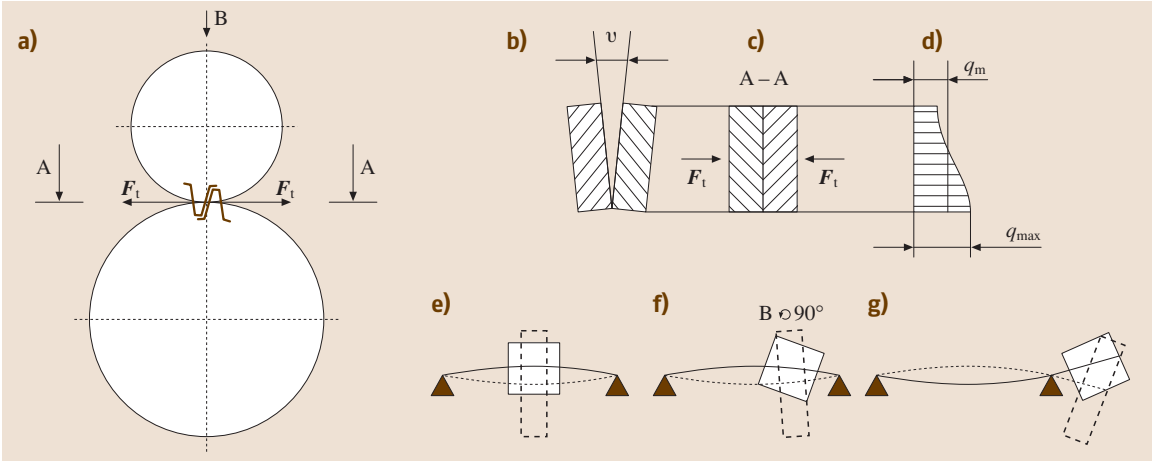


Fig. 6.25a–g Change of the nominal gradient angle β of the tooth due to **(a)**–**(d)** manufacturing errors, **(e)**–**(g)** resilience of the shaft to bending

produces a relative warp ϑ of the teeth in initial contact without load (Fig. 6.25a,b). As a consequence of yielding of the teeth and their distortion under the action of the force F_t in the toothing, the contact occurs along the whole length (Fig. 6.25c). However, resiliency along the tooth lengths are not equal, which results in an uneven load distribution, which is quantified by the ratio q_{max}/q_m (Fig. 6.25d). The relative teeth warp ϑ as a result of the elastic shaft flexure strain depends on the gearing position relative to the bearings (Fig. 6.25e–g). There is no teeth warp in a symmetrical arrangement; the greatest warp is observed in the case of the cantilever location (Fig. 6.25g). The directional tooth error is regulated by the degree of gearing accuracy according to the contact standards.

Torsional strain of the pinion body under the action of the torque produces tooth contortion, i. e., a change of direction of the tooth along the crown width. From the direction of the torque supply T_1 on the pinion end 1 (Fig. 6.26) the angle of torsion γ is maximum, but at end 2, $T = 0$, and torsional strains are absent, $\gamma = 0$. The degree of the change of direction of the tooth and the load distribution unevenness are higher for larger width b_2 of the gear. Moreover, how low the angular hardness c of the pinion is depends on the diameter d_1. This is why the factor $K_{H\beta}$ is chosen depending on the ratio $\psi_{bd} = b_2/d_1$.

The gear wheel teeth can *grind* and, as a result of increased local wear, the load distribution becomes even. Therefore, the unbalance factors are considered *at the initial time of functioning* $K^0_{H\beta}$ and after *grind* $K_{H\beta}$.

The value of the coefficient $K^0_{H\beta}$ is determined by the degree of accuracy required by the contact standards, the ratio $\psi_{bd} = b_2/d_1$, the layout of the gearing with regard to the bearings, and the teeth hardness ($K^0_{H\beta} = 1.05–1.5$). The value ψ_{bd} is computed from $\psi_{bd} = 0.5\psi_{ba}\ (u \pm 1)$, where ψ_{ba} is the gear face ratio.

The factor $K_{H\beta}$ is computed from

$$K_{H\beta} = 1 + \left(K^0_{H\beta} - 1\right) K_{Hw} ,$$

where K_{Hw} is a coefficient that takes into account teeth burn-in; its value is found in dependence on the circumferential velocity of the gear wheel with lower hardness. The grind capability decreases (K_{Hw} increase) with increasing hardness and circumferential velocity. The increase in the circumferential velocity assists in producing a steady oil layer between the teeth, which protects them from wear.

To lower the unbalance factor $K_{H\beta}$ along the contact line the wheels should be positioned symmetrically relative to the bearings, and the hardness of the gear wheels, shafts, and bearings should be increased (roller bearings should be applied instead of ball bearings), the manufacturing accuracy (of the gear wheels themselves, the bores for the bearings in the cases, etc.) should be increased, and barrel teeth should be used (Fig. 6.27).

The internal dynamic load in the toothing (the ratio K_{HV}) is caused by impact of the teeth on entry into the toothing due to production errors of the pitch and teeth distortion under loading. For nonimpact functioning, it is necessary that the teeth mesh and leave the toothing along the line of action, i. e., circular pitches must be equal under load. If the circular pitch is less than the wheel pitch, untimely meshing of the second teeth pair occurs and *edge impact* (at the tooth point of the driven wheel) is observed. When the circular pitch of the pinion is larger than the wheel pitch delay, the preceding teeth pair leaves the meshing, and thus the following pair does not make contact at the beginning of the impact, but rather in the middle of the working area of the contact line, a phenomenon known as *middle impact*.

Nominal force F in the toothing increases in F_{im} at impact. Then the full dynamic load is

$$F_{dy} = F + F_{im} = F(1 + F_{im}/F) = FK_{HV} .$$

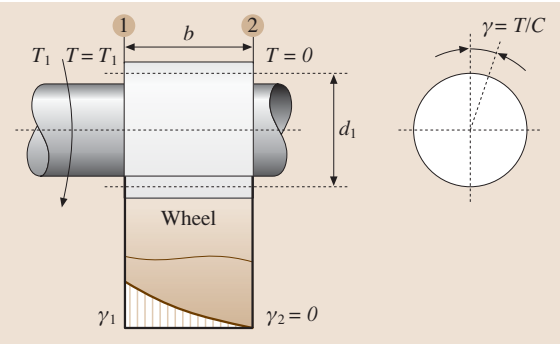


Fig. 6.26 Change of the tooth line as a consequence of torsional deformation of the pinion body

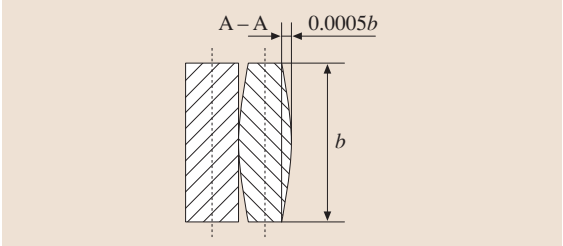


Fig. 6.27 Barrel-type form of the tooth

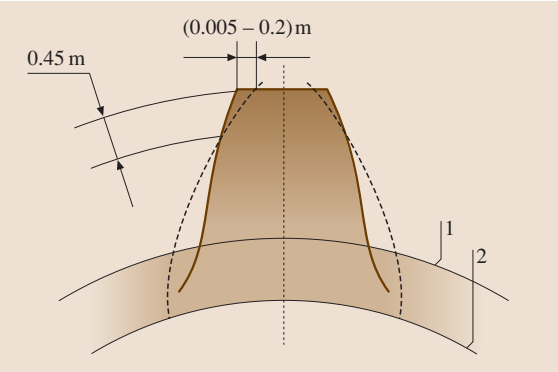


Fig. 6.28 Modification of the tooth point

From this formula it follows that higher surface hardness corresponds to a lower value of the ratio K_{HV}. It is caused by the fact that the nominal force is higher for greater hardness of the working teeth surfaces, whereas the impact force F_{im}, which depends on the degree of manufacturing accuracy and circumferential velocity, is the same.

The ratio of the internal dynamic load K_{HV} depends on the degree of gearing accuracy according to the smoothness standards, the circumferential velocity, and the hardness of the working surfaces ($K_{HV} = 1.01–1.6$). Lower values of K_{HV} correspond to higher-accuracy, helical, hard gearings that function with low circumferential velocity.

To reduce internal dynamic load production the accuracy should be increased according to the smoothness standards, helical or herring-bone gears should be used, and the tooth tip (teeth with a cut-off crest) should be modified. In order to obey the toothing theorem, cutting off is carried out according to the evolvent of the main circle 2 with diameter less than that of than the main circle 1 (Fig. 6.28).

The values of the coefficients $K^0_{H\beta}$, K_{HV}, and K_{Hw} are given as tables and charts [6.35].

The unevenness of the load distribution between the teeth (the ratio $K_{H\alpha}$) depends on production errors (pitch errors and teeth direction). Consequently, with the touching of one teeth pair of mating wheels, clearance is possible in another pair. Owing to teeth distortion on load application the clearance can disappear, but in this case load distribution unevenness is unavoidable: teeth with initial contact are more heavily loaded, while teeth with initial clearance are less heavily loaded. The ratio $K_{H\alpha}$ is determined by taking into account possible grinding as a result of increased local wear, which is why load distribution coefficients in the *initial operating period* $K^0_{H\alpha}$ and *after grinding* $K_{H\alpha}$ are considered

$$K_{H\alpha} = 1 + \left(K^0_{H\alpha} - 1\right) K_{Hw} ,$$

where K_{Hw} is a coefficient taking teeth grind into account.

In the bending stress calculation the load ratio is

$$K_F = K_A K_{F\beta} K_{FV} K_{F\alpha} ,$$

where K_A is the ratio of the *internal dynamic load* and is determined in the same way as in contact strength calculations. $K_{F\beta}$ is a ratio taking the *load distribution unevenness at the tooth root along the width of the gear ring* into consideration.

A lower influence of the load distribution unevenness on the bending stress ($K_{F\beta}$ is less than $K^0_{H\beta}$) is caused by bending stresses act throughout the tooth volume in contrast to contact stresses concentrated in the contact area.

K_{FV} is a ratio taking the *internal dynamic load* into account, and depends on the accuracy degree of the gearing according to the smoothness standards, circumferential velocity, and surface hardness of the wheel teeth ($K_{FV} = 1.01–2$). Lower values correspond to helical hard gearings of high accuracy that function with low circumferential velocities.

$K_{F\alpha}$ is a ratio taking *the influence of the production errors of the pinion and the wheel on the load distribution between the teeth* into account and is determined the same way as in contact strength calculations: $K_{F\alpha} = K^0_{H\alpha}$.

The influence of the grind on the bending strength is less favorable than it is on the contact strength. This can have severe consequences due to inaccuracy in the determination of bending stresses if the grind is not taken into consideration in the calculations of the ratios $K_{F\beta}$ and $K_{F\alpha}$.

6.3 Cylindrical Gearings

6.3.1 Toothing Forces of Cylindrical Gearings

It is customary to determine the interacting teeth forces at the pitch point. The load q is distributed throughout the contact area; in the toothing it changes into the resultant F_n normal to the tooth surface.

For the calculation of the shafts and bearings it is convenient to represent the force F_n by its components F_t, F_a, F_r (Fig. 6.29).

The circumferential force is

$$F_t = 2\times 10^3 T/d \,,$$

and the axial force is

$$F_a = F_t \tan\beta \,.$$

On the driven wheel the direction of the circumferential force F_t coincides with the rotational direction, whereas on the drive wheel the opposite is true. The axial force is parallel to the wheel axis. The direction of the vector F_a depends on the direction of wheel rotation and the tooth line.

For determination of the radial force F_r let us write the intermediate expression

$$F_R = F_t/\cos\beta \,.$$

Then the radial force (see the section A–A in Fig. 6.29)

$$F_r = F_R \tan\alpha_w = F_t \tan\alpha_w/\cos\beta \,,$$

where T is the torque of the gear wheel (N m), d is the pitch diameter of the wheel (in mm), β is the tilt angle of the tooth, and $\alpha_w = 20°$ is the angle of action. The radial force vectors of wheels with external toothing are directed towards the axis, whereas the vectors of wheels with internal toothing are directed away from the axis of the gear wheel.

6.3.2 Contact Strength Analysis of Straight Cylindrical Gearings

The contact strength of the teeth is the main efficiency specification for most gearings. In the derivation of the calculated relation according to the contact strength condition, teeth contact is considered at the pitch point W, i. e., in the zone of one-pair contact where flaking is observed. In Fig. 6.30 $O_1O_2 = a_w$ is an axle base, N_1N_2 is a contact line (the tangent to the base circles), α_w is an angle of action, d_{b1} and d_{b2} are the diameters of the base circles, and d_{w1} and d_{w2} are the diameters of the initial circles. In gearings without shifting pitch, the initial circles coincide: $d = d_w$.

As shown previously, we have

$$a_w = (d_2 \pm d_1)/2 = d_1(u \pm 1)/2 \,.$$

Hence,

$$d_1 = 2a_w/(u \pm 1) \quad \text{and} \quad d_2 = 2a_w u/(u \pm 1) \,,$$

where $u = d_2/d_1$ is the gear ratio of the gearing. The highest contact stress in the contact zone is determined from Hertz's formula (6.1) derived for the contact of two cylinders with parallel axes (Fig. 6.6).

To develop the calculated relation, let us express the values included in Hertz's formula using the toothing parameters. The force F_n acting along the normal to the profiles (along the toothing line in the contact point) is determined according to the circumferential force F_t taking into account the load ratio K_H

$$F_n = K_H F_t/\cos\alpha_w \,.$$

The length l_Σ of the contact lines in the toothing of the gear wheels with straight teeth changes from the effective gear face b_2 of the wheel in the zone of one-pair contact to $2b_2$ in the zone of two-pair contact. The rotation angle of the gear wheel by the transference of the profile tangent point from one extreme position to another is called the *front overlap angle*. The relation of the front overlap angle to the angle spacing $2\pi/z$

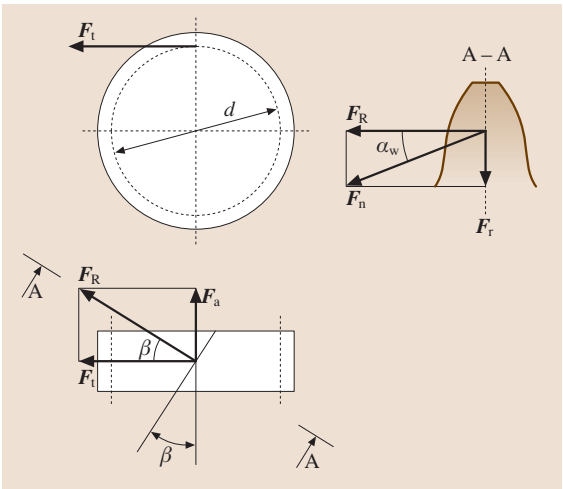


Fig. 6.29 Forces in the toothing of cylindrical gearings

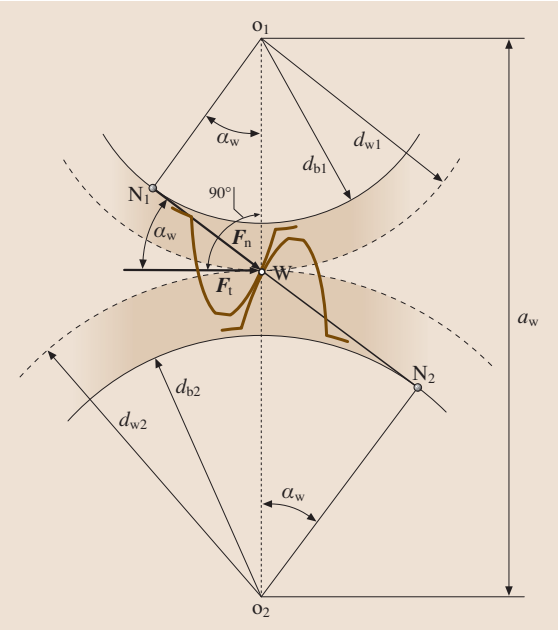


Fig. 6.30 Contact strength analysis of gearings

is called the *front contact ratio* ε_α. Here z is the teeth number of the gear wheel.

In compliance with the experimental results for the calculations the total length $b = l_\sum$ of the contact lines is determined, taking into account the front contact ratio ε_α, as

$$b = l_\sum = 3b_2/(4-\varepsilon_\alpha)\,.$$

Teeth contact is considered as the contact of two cylinders with radii equal to the curvature radii of the teeth profiles at the pitch point $\rho_1 = \mathrm{N_1W}$ and $\rho_2 = \mathrm{N_2W}$. Then

$$\sum(1/\rho_i) = 1/\rho_1 \pm 1/\rho_2 = (\rho_2 \pm \rho_1)\,/\,(\rho_1\rho_2)\,.$$

From the right triangle $\mathrm{O_1N_1W}$ we have: $\rho_1 = \mathrm{N_1W} = 0.5d_1\sin\alpha_\mathrm{w}$; and from the triangle $\mathrm{O_2N_2W}$ we have: $\rho_2 = \mathrm{N_2W} = 0.5d_2\sin\alpha_\mathrm{w} = 0.5ud_1\sin\alpha_\mathrm{w}$.

Then

$$\sum(1/\rho_i) = \frac{0.5ud_1\sin\alpha_\mathrm{w} \pm 0.5d_1\sin\alpha_\mathrm{w}}{0.5d_1\sin\alpha_\mathrm{w}0.5ud_1\sin\alpha_\mathrm{w}} = \frac{u\pm1}{0.5ud_1\sin\alpha_\mathrm{w}}\,.$$

Let us substitute the given relations into Hertz's formula

$$\sigma_\mathrm{H} = \sqrt{\frac{1}{\pi\left[\left(1-v_1^2\right)/E_1 + \left(1-v_2^2\right)/E_2\right]}} \times \sqrt{\frac{K_\mathrm{H}F_t}{\cos\alpha_\mathrm{w}}\,\frac{4-\varepsilon_\alpha}{3b_2}\,\frac{2}{d_1\sin\alpha_\mathrm{w}}\,\frac{u\pm1}{u}}\,.$$

Let us identify

$$Z_\mathrm{E} = \sqrt{1/\left[\pi[(1-v_1^2)/E_1 + (1-v_2^2)/E_2]\right]}$$

as the ratio that takes into consideration the stress–strain material properties of the mated wheels. For steel wheels $Z_\mathrm{E} = 191.6\,\mathrm{N/mm^2}$ for $E_1 = E_2 = 2.1\times10^5\,\mathrm{N/mm^2}$, and $v_1 = v_2 = 0.3$.
$Z_\mathrm{H} = \sqrt{2/\left(\cos\alpha_\mathrm{w}\sin\alpha_\mathrm{w}\right)}$ is the ratio that takes into account the form of the mated teeth surfaces at the pitch point; $Z_\mathrm{H} = 2.5$ for $\alpha_\mathrm{w} = 20°$. $Z_\varepsilon = \sqrt{(4-\varepsilon_\alpha)/3}$ is the ratio that takes into account the total length of the contact lines. $Z_\varepsilon = 0.9$ for spurs with $\varepsilon_\alpha = 1.6$.

Thus, we determine the relation in the form suggested by the standards

$$\sigma_\mathrm{H} = Z_\mathrm{E}Z_\mathrm{H}Z_\varepsilon\sqrt{\frac{K_\mathrm{H}F_\mathrm{t}\,(u\pm1)}{b_2d_1u}}\,, \qquad (6.4)$$

where u is the gear ratio; F_t is in N; d_1 and b_2 are in mm, and σ_H is in $\mathrm{N/mm^2}$.

Substituting into this formula $F_\mathrm{t} = 2\times10^3T_1/d_1$ and expressing b_2 and d_1 through a_w: $b_2 = \psi_{ba}a_\mathrm{w}$ and $d_1 = 2a_\mathrm{w}/(u\pm1)$, we then have

$$\sigma_\mathrm{H} = Z_\mathrm{E}Z_\mathrm{H}Z_\varepsilon\sqrt{\frac{K_\mathrm{H}2\times10^3T_1\,(u\pm1)}{b_2d_1^2u}} = Z_\mathrm{E}Z_\mathrm{H}Z_\varepsilon\sqrt{\frac{K_\mathrm{H}2\times10^3T_1\,(u\pm1)^3}{\psi_{ba}a_\mathrm{w}4a_\mathrm{w}^2u}}\,,$$

where T_1 is the torque of the pinion (N m).

We write the strength condition $\sigma_\mathrm{H} \le [\sigma]_\mathrm{H}$ in the form

$$\sigma_\mathrm{H} = Z_\mathrm{E}Z_\mathrm{H}Z_\varepsilon\sqrt{\frac{K_\mathrm{H}500T_1\,(u\pm1)^3}{\psi_{ba}a_\mathrm{w}^3u}} \le [\sigma]_\mathrm{H}\,. \qquad (6.5)$$

Having done this relative to a_w, we obtain

$$a_\mathrm{w} = (u\pm1)\sqrt[3]{500\,(Z_\mathrm{E}Z_\mathrm{H}Z_\varepsilon)^2}\sqrt[3]{\frac{K_\mathrm{H}T_1}{\psi_{ba}u\,[\sigma]_\mathrm{H}^2}}\,.$$

We designate

$$K_a = \sqrt[3]{500\,(Z_\mathrm{E}Z_\mathrm{H}Z_\varepsilon)^2}\,.$$

Finally the formula for the projection analysis of cylindrical gearings takes the following form

$$a_w = K_a\,(u \pm 1)\,\sqrt[3]{\frac{K_H T_1}{\psi_{ba} u\,[\sigma]_H^2}}\,,$$

where a_w is an axle base (in mm), K_H is a load ratio, T_1 is a torque on the pinion (N m), and $[\sigma]_H$ is an allowable contact stress (in units of N/mm^2).

In accordance with the standards:

- For straight gearings $K_a = 450\,(\mathrm{N/mm^2})^{1/3}$
- For helical and herring-bone gearings $K_a = 410\,(\mathrm{N/mm^2})^{1/3}$

In general, the axle base of helical cylindrical gearings is approximately 20% less than that of straight gearings.

On the analysis of cylindrical gearings the ratio value of the wheel face width $\psi_{ba} = b_2/a_w$ is set. *Depending on the pinion position relative to the bearings* it is taken as $\psi_{ba} = 0.2$–0.5.

On the basis of (6.5) we obtain the formula for the verification calculation

$$\sigma_H = \frac{Z_E Z_H Z_\varepsilon \sqrt{500}}{a_w}\sqrt{\frac{K_H T_1\,(u \pm 1)^3}{\psi_{ba} a_w u}}\,.$$

Using $Z_\sigma = Z_E Z_H Z_\varepsilon \sqrt{500}$ and substituting in $\psi_{ba} a_w = b_2$, we obtain the formula for the verification analysis of cylindrical gearings

$$\sigma_H = \frac{Z_\sigma}{a_w}\sqrt{\frac{K_H T_1\,(u \pm 1)^3}{b_2 u}} \le [\sigma]_H\,, \qquad (6.6)$$

where T_1 is in N m, a_w and b_2 are in mm, and σ_H is in N/mm^2.

The values of the ratio Z_σ for cylindrical steel gearings are:

- For straight gearings $Z_\sigma = 9600\,(\mathrm{N/mm^2})^{1/2}$
- For helical and herring-bone gearings $Z_\sigma = 8400\,(\mathrm{N/mm^2})^{1/2}$

In the projection analysis the ratio value of the rated load is set approximately as $K_H = 1.3$. In the checking analysis its adjusted value is determined from all known gearing parameters.

From the given formulas it follows that the contact strength of the wheel teeth depends on the material and gearing dimensions, and does not depend on the module and teeth number individually. With a contact strength with given a_w the module and teeth number can have different values if they meet the conditions $0.5m(z_1 + z_2) = a_w$ and $u = z_2/z_1$.

6.3.3 Bending Strength Calculation of Cylindrical Gearing Teeth

The second of the two basic efficiency criteria for gearings is the bending strength of the teeth. For the derivation of the rating relation it is supposed that (Fig. 6.31):

1. There is one pair of teeth in contact.
2. The tooth is considered as a cantilever bar loaded with a concentrated force F_n applied to the tooth point.

The force F_n acts at an angle $(90° - \alpha')$ to the symmetry axis of the tooth. The angle α' is slightly larger than the angle of action α_w. For the determination of the tooth stress state the force F_n is transferred along the contact line N_1N_2 to the intersection with the tooth axis at point C (Fig. 6.32a), decomposed into the constituents directed along the tooth axis and transversely to it.

Under the influence of the constituent directed along the axis, compressions act at the tooth root $\sigma_{com} = F_n \sin\alpha'/(bS)$, where b is the tooth length (Fig. 6.32b).

The points A and B determine the position of the weak tooth section in the case of bending. The

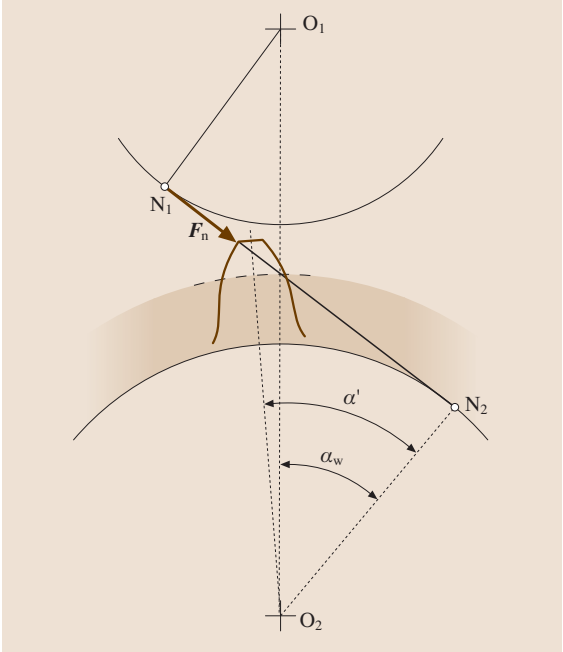


Fig. 6.31 Bending strength analysis of teeth

tooth in this section is loaded with bending moment $M = F_n h_c \cos\alpha'$, producing stress action σ_b. On the left-hand side of the axis in Fig. 6.32b there is tension; on the right-hand side there is compression.

Total stresses $\sigma_{F\,nom}$ from the side of the tension (point A) have lower values than from the side of the compression (point B). However, tensile stresses are more dangerous. From experience in the field the fatigue crack 1 that results in breaking off of the tooth arises exactly from the side of the tension at point A (Fig. 6.32). Stresses found without taking into account stress concentrators are called *nominal*.

Let us determine the nominal stresses $\sigma_{F\,nom}$ of the bending compression at point A

$$\sigma_{F\,nom} = \sigma_b - \sigma_{com} = \frac{M}{W_x} - \frac{F_n \sin\alpha'}{bS}$$
$$= \frac{F_n \cos\alpha' h_c 6}{bS^2} - \frac{F_n \sin\alpha'}{bS}$$
$$= \frac{F_n}{b}\left(\frac{\cos\alpha' h_c 6}{S^2} - \frac{\sin\alpha'}{S}\right),$$

where $W_x = bS^2/6$ is the axial modulus of the weak section AB.

Expressing the force F_n through the circumferential force F_t, and taking the load ratio K_F: $F_n = K_F F_t/\cos\alpha_w$ into account, one finds

$$\sigma_{F\,nom} = \frac{K_F F_t}{b}\frac{1}{\cos\alpha_w}\left(\frac{\cos\alpha' h_c 6}{S^2} - \frac{\sin\alpha'}{S}\right).$$

The weak section AB is positioned in the zone of stress concentration that is caused by form change on the fillet surface at the tooth root. Local stresses in this section increase nominal stresses α_T by a factor of

$$\sigma_F = \sigma_{F\,nom}\alpha_T,$$

where α_T is the theoretical stress concentration factor.

Taking into consideration the stresses in the weak section, one obtains

$$\sigma_F = \frac{K_F F_t}{b}\frac{1}{\cos\alpha_w}\left(\frac{\cos\alpha' h_c 6}{S^2} - \frac{\sin\alpha'}{S}\right)\alpha_T.$$

Value h_c and tooth thickness S are expressed through the normal module m

$$h_c = \mu m \quad \text{and} \quad S = \lambda m,$$

where μ and λ are factors taking the tooth shape into account.

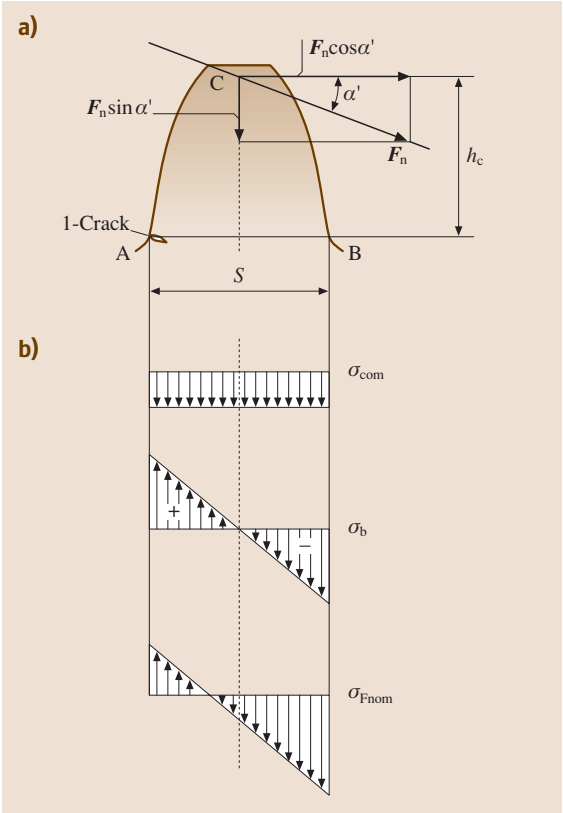


Fig. 6.32a,b Stressed state of the tooth. **(a)** Loading diagram, **(b)** stress distribution diagrams

Then

$$\sigma_F = \frac{K_F F_t}{b}\frac{1}{\cos\alpha_w}\left(\frac{\cos\alpha'\mu m 6}{\lambda^2 m^2} - \frac{\sin\alpha'}{\lambda m}\right)\alpha_T$$
$$= \frac{K_F F_t}{bm}\frac{1}{\cos\alpha_w}\left(\frac{\cos\alpha'\mu 6}{\lambda^2} - \frac{\sin\alpha'}{\lambda}\right)\alpha_T$$
$$= \frac{K_F F_t}{bm} Y_{FS},$$

where Y_{FS} is the ratio taking into consideration tooth shape and stress concentration

$$Y_{FS} = \frac{1}{\cos\alpha_w}\left(\frac{\cos\alpha'\mu 6}{\lambda^2} - \frac{\sin\alpha'}{\lambda}\right)\alpha_T.$$

The values of the ratio $Y_{FS} = 4.5–3.6$ taking tooth form and stress concentration into consideration are given in the form of tables and charts [6.35]. Lower values of the ratio Y_{FS} correspond to larger numbers of teeth and a positive shift of the tools because both result in an increase of the tooth thickness at the root.

Considering the strength condition $\sigma_F \le [\sigma]_F$, we obtain the formula for verification of the bending stress analysis of the gearings

$$\sigma_F = \frac{K_F F_t}{bm} Y_{FS} Y_\beta Y_\varepsilon \le [\sigma]_F \ , \qquad (6.7)$$

where $[\sigma]_F$ is an allowable bend stress in N/mm^2, F_t is in N, and b and m are in mm.

In the given formula, we additionally have that Y_β is the ratio that takes the tilt angle of the tooth into account, and Y_ε is the ratio that takes teeth overlap into consideration.

For spurs, $Y_\beta = 1$, $Y_\varepsilon = 1$, with an accuracy degree of 8 and 9, and $Y_\varepsilon = 0.8$, with an accuracy degree of 5–7.

Owing to the smaller number of teeth on the pinion, their roots are thinner than those on the wheel, which is reflected in a higher value of the ratio Y_{FS} (i. e., $Y_{FS1} > Y_{FS2}$). To provide an approximately equal bending strength of the pinion teeth and the wheel teeth, the pinion is manufactured from more resistant material compared with that used for the wheel.

The condition for equal bending strength for teeth on the and the wheel is

$$[\sigma]_{F1}/Y_{FS1} \approx [\sigma]_{F2}/Y_{FS2} \ .$$

Substituting into (6.7) $F_t = 2\times10^3 T_1/d_1$ and $d_1 = 2a_w/(u \pm 1)$ we obtain the formula for checking the bending stress analysis of the teeth

$$\sigma_F = \frac{K_F 10^3 T_1 \ (u \pm 1)}{bma_w} Y_{FS} Y_\beta Y_\varepsilon \le [\sigma]_F \ ,$$

where T_1 is in N m; b_2, m and a_w are in mm, and σ_F and $[\sigma]_F$ are in N/mm^2.

This inequality can be rearranged in terms of m as

$$m \ge \frac{K_F T_1 \ (u \pm 1)}{ba_w \ [\sigma]_F} 10^3 Y_{FS} Y_\beta Y_\varepsilon \ .$$

The pinion face width b_1 is 2–4 mm thicker than the wheel width b_2 to balance possible axial shifting of the gear wheels due to installation inaccuracy. This condition is important for teeth grind, when a harder pinion overlaps edgewise with a less hard wheel.

Supposing that $b = b_2$ and $K_m = 10^3 Y_{FS} Y_\beta Y_\varepsilon$ we obtain the rated relation for the determination of the minimum value of the teeth module

$$m \ge K_m K_F T_1 (u \pm 1)/(b_2 a_w [\sigma]_F) \ ,$$

where $K_m = 3.4\times10^3$ for straight gearings and $K_m = 2.8\times10^3$ for helical gearings, and T_1 is in N m, b_2, a_w are in mm, and $[\sigma]_F$ is in N/mm^2. Instead of $[\sigma]_F$ the lowest of $[\sigma]_{F1}$ and $[\sigma]_{F2}$ is substituted into this formula.

6.3.4 Geometry and Working Condition Features of Helical Gearings

The teeth of helical cylindrical wheels are cut with the same tool as is used for spurs. The hob axis forms an angle β with the end face of the wheel (Fig. 6.33). During cutting, the hob is moved in the direction of the wheel teeth. Thus, all dimensions are the same as in the straight cylindrical gearings in the plane normal to the tooth surface.

A pair of mating helical wheels with external toothing have equal angles β, but in opposite directions. In the absence of special demands, wheels are cut in the right tooth direction, and the pinions in the left direction.

The distance between the teeth of the helical wheel (Fig. 6.33) can be measured in the end, peripheral, (t–t), and standard (n–n) directions. In the first case the peripheral pitch p_t is obtained, in the second case the standard pitch p is obtained. The toothing modules are also different in these directions

$$m_t = p_t/\pi; \quad m_n = p/\pi \ ,$$

where m_t and m_n are the peripheral and normal modules of the teeth.

In compliance with Fig. 6.33

$$p_t = p/\cos\beta \ ,$$

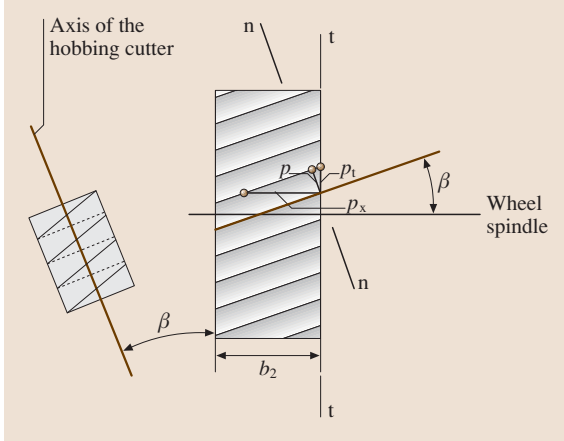


Fig. 6.33 Main parameters of a cylindrical helical wheel

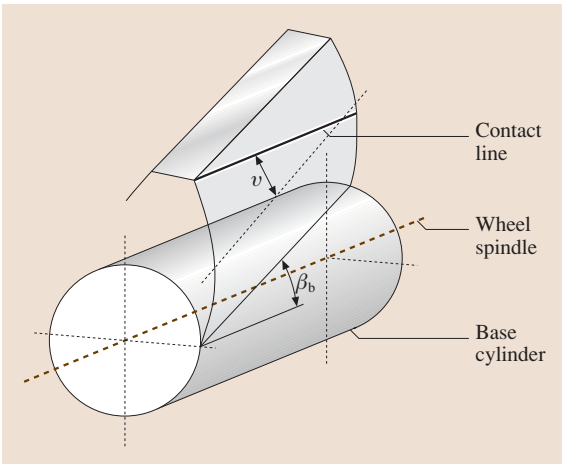


Fig. 6.34 Position of the contact line on a skew tooth of the wheel

and therefore

$$m_t = m/\cos\beta\ ,$$

where β is the tilt angle of the tooth on the pitch cylinder. The standard module conforms to the standard.

On the end face the t–t helical wheel can be considered as a spur with module value m_t, and angle of action α_T: $\tan\alpha_T = \tan\alpha/\cos\beta$.

For a wheel without a shifting pitch d and starting d_w diameters

$$d = d_w = m_t z = m_n z/\cos\beta\ .$$

In addition to front overlap there is also axial overlap in helical gearings. The axial overlap factor is

$$\varepsilon_\beta = b_2/p_x\ ,$$

where p_x is an axial spacing that equals the distance between the analogous points of two adjacent teeth, measured in the direction of the gear wheel axis (Fig. 6.33)

$$p_x = \pi m_n/\sin\beta\ .$$

Differences in the Functioning Conditions of Helical Gearings Determine Their Geometry Features

1. The contact lines on the helical wheel are positioned parallel to the axis of revolution (Fig. 6.34), at an angle of ϑ to the pitch line (on the spur this is parallel to the pitch line). Here β_b is the tilt angle of the tooth on the basic cylinder. The driven wheel tooth meshes starting from the tooth point, first increasing, then reducing the contact line length when traversing it from the tooth tip to the dedendum. As a result of the fact that the tooth does not operate along the whole length at once, it grinds better and more quickly.
2. In contrast to spur gears, in helical gearings the teeth do not mesh along the whole length at once, but rather gradually. The contact time of one teeth pair increases while new teeth pairs interlock; more contact lines transmit the load, which reduces noise and dynamic loads considerably. The greater the tilt angle β of the tooth line (Fig. 6.33), the greater the smoothness of the toothing.
3. The load along the contact line is distributed proportionally to the total teeth hardness of the pinion and the wheel (Fig. 6.35a). The contact of the mating teeth in their typical sections and their schematic sketch to determine the total rigidity are shown in Fig. 6.35b. The hardness is lower and the load is lower for the contact of one of the mating teeth at

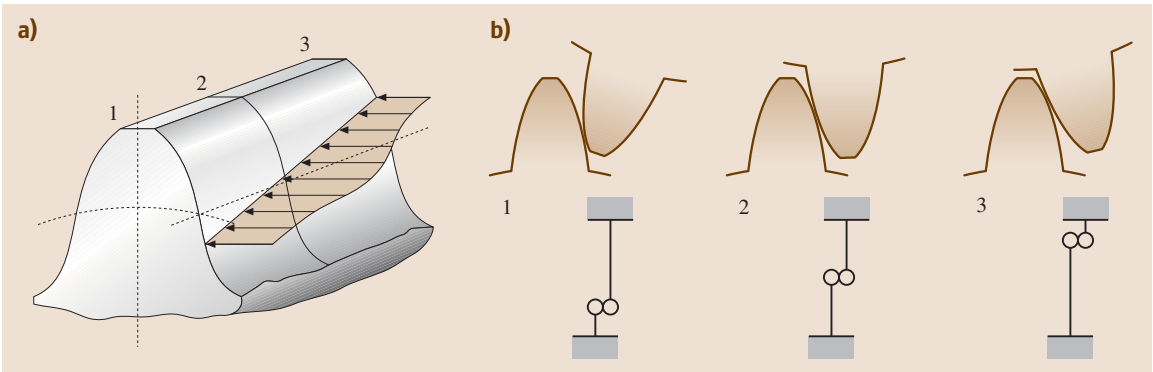


Fig. 6.35a,b Contact of conjugated teeth in the typical sections **(a)** and schematic sketch **(b)**

a point (profiles 1 and 3). Such a load distribution is positive for the gearing operation.

4. Due to axial overlap in helical gearings two or three pairs of the teeth take part in toothing simultaneously. Thus, the total length $l_{\sum \text{hel}}$ of the contact lines in the helical gearing is longer than in the spur $l_{\sum \text{spur}}$

$$l_{\sum \text{hel}} = \frac{b_2}{Z_\varepsilon^2 \cos\beta_b} \; ; \quad l_{\sum \text{spur}} = \frac{b_2}{Z_\varepsilon^2} \; .$$

In this regards, the values of the ratio Z_ε, taking into account the total length of the contact lines becomes

$$Z_\varepsilon = \sqrt{1/\varepsilon_\alpha} \quad \text{for helical gearings} \; ,$$
$$Z_\varepsilon = \sqrt{(4-\varepsilon_\alpha)/3} \quad \text{and for spur gears} \; ,$$

where ε_α is the front contact ratio.

5. The correlation between the radii of curvature of contacting teeth in the helical gearing is more favorable

$$\sum (1/\rho_i)_{\text{hel}} = \cos\beta_b \sum (1/\rho_i)_{\text{spur}} \; .$$

This is embodied in the calculation of the ratio Z_H, taking the form of the mated teeth surfaces into consideration. Contact stresses under other equal conditions in helical gearings are lower in value than in spur gearings.

6. The tooth form provides higher bend strength.

6.3.5 The Concept of the Equivalent Wheel

As has been mentioned, the profile of the oblique tooth in the standard section n–n (Fig. 6.33) coincides with the profile of the spur. The calculation for helical wheels is carried out using the parameters of the equivalent spur: the module m_n and the teeth number z_v.

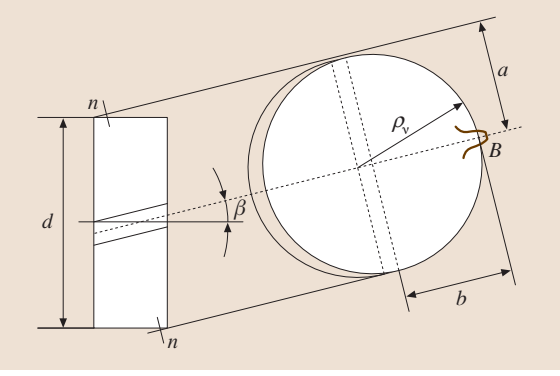


Fig. 6.36 Reduction of the helical wheel to the equivalent spur

The pitch cylinder of the helical wheel in the plane n–n normal to the tooth line (Fig. 6.36) forms an ellipse with the semi-axles: large $a = d/(2\cos\beta)$ and small $b = d/2$. The radius of curvature in the vertex B of the ellipse is

$$\rho_v = \frac{a^2}{b} = \frac{d^2}{4\cos^2\beta}\frac{2}{d} = \frac{d}{2\cos^2\beta} \; .$$

The tooth profile in this section coincides with the profile of the relative spur called the *equivalent*, the pitch diameter of which is $d_v = m_n z_v$.

Taking into consideration that $d = m_t z$ and $m_t = m_n/\cos\beta$ we have

$$d_v = 2\rho_v = d/\cos^2\beta = m_t z/\cos^2\beta = m_n z/\cos^3\beta \; .$$

From the equality

$$m_n z_v = m_n z/\cos^3\beta \; ,$$

the equivalent teeth number follows as

$$z_v = z/\cos^3\beta \; ,$$

where z is the actual teeth number of the helical wheel. With an increase in the tilt angle β of the tooth line, the equivalent parameters increase, which in turns leads to an increase in the strength of the gearings.

6.3.6 Strength Analysis Features of Helical Gearings

Strength analysis of helical gearings is carried out with calculation formulas of spur gears with insertion of correction factors that take into account their working features. These include high working smoothness (lower values of the factor of internal dynamic load K_v), longer total length of contact lines, more favorable combination of the curvature radii, and higher fatigue bend resistance (lower value of the factor Y_{FS} of the tooth form and stress concentrations as $z_v > z$). According to these strength conditions the dimensions of helical gearings are smaller than those of spur gears. In contact strength calculations the features of the geometry and work conditions of helical gearings are taken into consideration by means of the factors Z_H, Z_ε and K_H. The features of helical gearings in *bending strength testing* of the pinion teeth and wheel teeth are taken into account by the factors K_F, Y_{FS}, Y_β, and Y_ε. The factor Y_{FS} of the tooth form and stress concentrations is chosen according to the equivalent teeth number z_v. The

Table 6.1 Mechanical characteristics of steels used for the manufacture of gear wheels. Induction hardening is surface quenching with heating by means of a high-frequency current

Steel grade		Heat treatment	Limit uncut dimensions (mm)		Teeth hardness		σ_y (N/mm^2)
			D_{max}	S_{max}	In the core	On the surface	
45	C45 (EN)	Refining	125	80	235–262 HB	235–262 HB	540
		Refining	80	50	269–302 HB	269–302 HB	650
40X	41Cr4 (EN)	Refining	200	125	235–262 HB	235–262 HB	640
		Refining	125	80	269–302 HB	269–302 HB	750
		Refining and induction hardening	125	80	269–302 HB	45–58 HRC	750
40XH,	40NiCr6 (DIN)	Refining	315	200	235–262 HB	235–262 HB	630
35XM	34CrMo (EN)	Refining	200	125	269–302 HB	269–302 HB	750
		Refining and induction hardening	200	125	269–302 HB	48–55 HRC	750
40XH2MA, 38X2MIOA	36CrNiMo4 (EN) 41CrAlMo7 (DIN)	Refining and nitriding	125	80	269–302 HB	58–67 HRC	780
20X, 20XH2M, 18XГT, 12XH3A, 25XГM, 30XГT	20CrS4 (DIN) 17CrNiMo (DIN) 20MnCr5G (DIN) 14NiCr10 (DIN) 20CrMo5 (DIN) 30MnCrTi (DIN)	Refining, cementing (nitrocementing) and quenching	200	125	300–400 HB	56–63 HRC	800

factor Y_β, which takes into account the tooth tilt in helical gearings, is determined from the following formula (where β is in degrees)

$$Y_\beta = 1 - \varepsilon_\beta \beta / 120 \,,$$

under the condition that $Y_\beta \geq 0.7$.

The factor Y_ε takes into consideration teeth overlap in helical gearings, where $Y_\varepsilon = 1/\varepsilon_\alpha \approx 1/1.6 \approx 0.65$.

6.3.7 The Projection Calculation of Cylindrical Gearings

The basic data for the calculation under typical loading conditions are the the turning moment T_1 (N m), the rotational frequency of the pinion rotation n_1 (min^{-1}), the gear ratio u, the working time of the gearing (lifetime) L_h (h), and the layout drawing and type of the gearing.

1. Choosing of the wheel hardness, thermal treatment and material

   For power trains the most frequently used material is steel (Table 6.1). Below the steels used for gearing production in Russia are described (Appendix 6.A Table 6.95).

   In the case of surface heat or chemicothermal teeth treatment, the previous heat refining treatment determines the mechanical data of the tooth core.

2. Allowable contact stresses $[\sigma]_{H1}$ for the pinion and $[\sigma]_{H2}$ for the wheel are determined from the total relation, but with substitution of the appropriate parameters for the pinion and the wheel

$$[\sigma]_H = \sigma_{H\,lim} Z_N Z_R Z_V / S_H \,.$$

The fatigue contact point $\sigma_{H\,lim}$ is determined from empirical formulas depending on the material and heat treatment method of the gear wheel and the average hardness (HB_m or HRC_m) of the teeth surface (Table 6.2).

The minimum value of the load factor for gear wheels with homogenous material structure (refined, volume-quenched) is $S_H = 1.1$; for gear wheels with case-hardening one has $S_H = 1.2$.

The service life ratio Z_N takes into account the influence of lifetime:

for $N_{HE} \leq N_{H\,lim}$

$$Z_N = \sqrt[6]{N_{H\,lim}/N_{HE}} \,,$$

Table 6.2 Contact fatigue points $\sigma_{H\,lim}$

Method of heat or chemicothermal treatment	Average hardness on the surface	Steel	$\sigma_{H\,lim}$ (N/mm^2)
Refining, normalization	< 350 HB	Carbon and alloy steel	2 HB_m + 70
Surface and volume hardening	40–56 HRC		17 HRC_m + 200
Cementing, nitrocementing	> 56 HRC	Alloy steel	23 HRC_m
Nitriding	> 58 HRC		1050

under the condition that $Z_N \le Z_{N\,max}$;
for $N_{HE} > N_{H\,lim}$

$$Z_N = \sqrt[20]{N_{H\,lim}/N_{HE}}\,,$$

under the condition that $Z_N \ge 0.8$.

The number $N_{H\,lim}$ of cycles corresponding to the change of the stress-cycle diagram is determined from the average surface hardness of the teeth

$$N_{H\,lim} = 30\,HB_m^{2.4} \le 12\times10^7\,.$$

Hardness in units of HRC is transferred into units HB according to Table 6.3.

Table 6.3 Hardness in units of HRC corresponding to units HB

HRC	HB
45	427
47	451
48	461
50	484
51	496
53	521
55	545
60	611
62	641
65	688

In fatigue resistance calculations the action of the short-term overload moment T_{max} is not taken into consideration, but actual varying loading conditions are changed with equivalent constant conditions with nominal torque T and equivalent number N_{HE} of the loading cycles

$$N_{HE} = \mu_H N_k\,.$$

The values of the equivalence factor μ_H for typical loading conditions are given in Table 6.4.
The lifetime N_k of the gearing in cycle numbers of stress changes with rotational frequency n (min^{-1}) and working time L_h (h) is

$$N_k = 60 n n_t L_h\,,$$

where n_t is the number of tooth matings of the calculated wheel per revolution. Generally the total working lifetime L_h (h) of the gearing is determined from

$$L_h = L365K_{year}24K_{day}\,,$$

where L is the number of working years; K_{year} and K_{day} are, respectively, ratios of annual and daily gearing applications.
If plastic strain and brittle fracture of the surface layer are prevented, $Z_{N\,max} = 2.6$ for materials with homogenous structure (refined, volume-quenched) and $Z_{N\,max} = 1.8$ for case-hardened materials (quenching by means of high-frequency current, cementing, nitrocementing, and nitriding).
The factor Z_R, which takes into account the *roughness influence* of the mated teeth surfaces, is specified for the gear wheel pair with a rougher surface depending on the parameter Ra of roughness ($Z_R = 1$–0.9). Higher values of this parameter correspond to grinding finishes and polished faces ($Ra = 0.63$–$1.25\,\mu m$).

The factor Z_V takes the circumferential velocity v influence into consideration

$$Z_V = 0.85v^{0.1} \ge 1, \qquad \text{by } H \le 350\,HB\,;$$
$$Z_V = 0.925v^{0.05} \ge 1, \qquad \text{by } H > 350\,HB\,.$$

The allowable stress $[\sigma]_H$ for cylindrical gearings with straight teeth is the lowest of the allowable stresses of the pinion $[\sigma]_{H1}$ and the wheel $[\sigma]_{H2}$.

Table 6.4 Coefficients of equivalence for typical loading conditions of gearings

Designation of the typical rate	Factors of equivalence μ_H	μ_F $q = 6$	$q = 9$
0	1	1.0	1.0
I	0.500	0.300	0.200
II	0.250	0.143	0.100
III	0.180	0.065	0.036
IV	0.125	0.038	0.016
V	0.063	0.013	0.004

Table 6.5 Fatigue points $\sigma_{F\,lim}$ with the zero-to-tension loading cycle of Russian steels (for international steel grades see Appendix 6.A.1)

The method of heat and chemicothermal treatment	Steel grade		Teeth hardness		$\sigma_{F\,lim}$ (N/mm^2)
			On the surface	In the core	
Refining	45, 40X, 40XH, 35XM	C45 (EN), 41Cr4 (EN), 40NiCr6 (DIN), 34CrMo (EN)	< 350 HB	< 350 HB	1.75 HB_m
Induction hardening along the teeth contour	40X, 40XH, 35XM	41Cr4 (EN), 40NiCr6 (DIN), 34CrMo (EN)	48–58 HRC	25–35 HRC	600–700
Through induction hardening ($m < 3$ mm)			48–55 HRC	48–55 HRC	500–600
Cementing	20X, 18XГT, 25XГP, 12XH3A, 20XH2M	20CrS4 (DIN), 20MnCr5G (DIN), 20CrMo5 (DIN), 14NiCr10 (DIN), 17CrNiMo (DIN)	56–63 HRC	30–45 HRC	750–800
Cementing with automatic control of the process					850–900
Nitrocementing	25XГM, 30XГT	20CrMo5 (DIN), 30MnCrTi (DIN)	56–63 HRC	30–45 HRC	650–850
Nitriding	38X2MЮA, 40XH2MA	41CrAlMo7 (DIN), 36CrNiMo4 (DIN)	58–67 HRC	24–40 HRC	11HRC_m^{core} +200

For cylindrical gearings with indirect teeth, because the contact line is at an angle to the pitch line, allowable stresses can be increased up to the value of

$$[\sigma]_H = \sqrt{0.5\left([\sigma]_{H1}^2 + [\sigma]_{H2}^2\right)}\,,$$

under the operating condition $[\sigma]_H \le 1.2[\sigma]_{H\,min}$, where $[\sigma]_{H\,min}$ is the lowest of $[\sigma]_{H1}$ and $[\sigma]_{H2}$.

The allowable stress for bevel gearings with straight and indirect teeth is the lowest of allowable stresses of the pinion $[\sigma]_{H1}$ and the wheel $[\sigma]_{H2}$.

3. The allowable bending stress of the teeth of the pinion $[\sigma]_{F1}$ and the wheel $[\sigma]_{F2}$ is determined from the total relation, but with substitution of the appropriate parameters for the pinion and the wheel, respectively

$$[\sigma]_F = \sigma_{F\,lim} Y_N Y_R Y_A / S_F\,.$$

The fatigue point $\sigma_{F\,lim}$ with the zero-to-tension stress cycle is taken according to Table 6.5.

The minimum values of the load factors for cemented and nitrocemented gear wheels are $S_F = 1.55$ and for the rest of the gear wheels $S_F = 1.7$. The service life ratio Y_N considers the lifetime influence

$$Y_N = \sqrt[q]{N_{F\,lim}/N_{FE}}\,,$$

under the condition that $1 \le Y_N \le Y_{N\,max}$, where $Y_{N\,max} = 4$ and $q = 6$ for refined gear wheels, and $Y_{N\,max} = 2.5$ and $q = 9$ for quenched and case-hardened teeth. The cycle number corresponding to the change of the stress-cycle diagram $N_{F\,lim} = 4 \times 10^6$.

The equivalent number of loading cycles is given by

$$N_{FE} = \mu_F N_k\,.$$

The values of the equivalence factor μ_F for typical loading conditions are given in Table 6.4. The assigned life N_k is calculated just as in contact stress analysis.

The factor Y_R considers the roughness influence of the tooth space and is taken to be $Y_R = 1$ in the case of grinding, and gear milling with roughness parameter $Rz \le 40\,\mu m$; whereas $Y_R = 1.05–1.2$ in

Table 6.6 Allowable circumferential velocities of gearings

Accuracy degree according to [6.28]	Allowable circumferential velocity v (m/s) of the wheels, no more than			
	Spur		Indirect	
	Cylindrical	Bevel	Cylindrical	Bevel
6 (gearings of extra accuracy)	20	12	30	30
7 (gearings of normal accuracy)	12	8	20	10
8 (gearings of reduced accuracy)	6	4	10	7
9 (gearings of low accuracy)	2	1.5	4	3

the case of polishing (large values are obtained by refining and after quenching with heating by means of high-frequency currents).
The factor Y_A takes the influence of double-sided load application (reversing gears) into account. For one-sided load application $Y_A = 1$. For reverse loading and an equal load and number of loading cycles in the forward and backward direction (e.g., the teeth of the satellite in planetary gearing) $Y_A = 0.65$ for normalized and refined steels, $Y_A = 0.75$ for hardened and cemented steel, and $Y_A = 0.9$ for nitrided steel.

4. The tentative value of the axle base is a'_w (mm)

$$a'_w = K\,(u \pm 1)\,\sqrt[3]{T_1/u}\,,$$

where the plus sign applies for external toothing, and the minus sign applies for internal toothing. T_1 is the torque on the pinion (the highest of long-acting), in N m, and u is the gear ratio.
The factor K, which depends on the surface hardnesses H_1 and H_2 of the teeth of the pinion and the wheel, respectively, takes the following values:

Table 6.7 Coefficient K for the cylindrical gearings

Hardness H		Factor K
$H_1 \le 350$ HB	$H_2 \le 350$ HB	10
$H_1 \ge 45$ HRC	$H_2 \le 350$ HB	8
$H_1 \ge 45$ HRC	$H_2 \ge 45$ HRC	6

The circumferential velocity v (m/s) is determined from the formula

$$v = \frac{2\pi a'_w n_1}{6\times 10^4\,(u \pm 1)}\,.$$

The accuracy degree is taken from Table 6.6.
The previously determined value of the axle base is specified according to the formula

$$a_w = K_a\,(u \pm 1)\,\sqrt[3]{\frac{K_H T_1}{\psi_{ba} u [\sigma]_H^2}}\,,$$

where $K_a = 450$ applies for spurs and $K_a = 410$ applies for helical and herring-bone gears $(\mathrm{N/mm^2})^{1/3}$; and $[\sigma]_H$ is in $\mathrm{N/mm^2}$. ψ_{ba} is a width ratio taken from the sequence of standard numbers: 0.1, 0.125, 0.16, 0.2, 0.25, 0.315, 0.4, 0.5, and 0.63 depending on the wheel position relative to the bearings. Its value is as follows

$$\begin{cases} 0.315\text{–}0.5 & \text{symmetrical arrangement} \\ 0.25\text{–}0.4 & \text{unsymmetrical arrangement} \\ 0.2\text{–}0.25 & \text{console arrangement of one or both wheels} \end{cases}.$$

For herring-bone gearings $\psi_{ba} = 0.4\text{–}0.63$, for gear-boxes $\psi_{ba} = 0.1\text{–}0.2$, and for gearings of internal toothing $\psi_{ba} = 0.2(u+1)/(u-1)$. Lower values ψ_{ba} are obtained for gearings with teeth hardness $H \ge 45$ HRC.

The load factor in contact strength calculations is

$$K_H = K_{HV} K_{H\beta} K_{H\alpha}\,.$$

The factor K_{HV} takes the internal dynamics of the loading into consideration. The values K_{HV} are taken from Table 6.8 and depend on the accuracy degree of the gearing according to the smoothness standards, the circumferential velocity, and the hardness of the working surfaces.
The factor $K_{H\beta}$ takes the unevenness of the load distribution along the length of the contact lines into account. The teeth of the gear wheels can grind, and thus the unbalance factors are considered during *the initial working period* $K^0_{H\beta}$ and *after grinding* $K_{H\beta}$.
Values of the factor $K^0_{H\beta}$ are taken from Table 6.9 and depend on the coefficient $\psi_{bd} = b_2/d_1$, the gearing layout, and the teeth hardness. As the wheel width and pinion diameter have not yet been determined, the value of the coefficient ψ_{bd} is calculated approximately as

$$\psi_{bd} = 0.5\psi_{ba}(u \pm 1)\,.$$

The factor $K_{H\beta}$ is determined from the formula

$$K_{H\beta} = 1 + \left(K^0_{H\beta} - 1\right) K_{Hw}\,,$$

where K_{Hw} is the factor taking into account teeth grinding; its values are computed depending on the circumferential velocity of the gear wheel with lower hardness (Table 6.10).
The factor $K_{H\alpha}$, which takes into consideration the load distribution between the teeth, is determined from the formula

$$K_{H\alpha} = 1 + \left(K^0_{H\alpha} - 1\right) K_{Hw}\,,$$

where K_{Hw} is the factor considering teeth grinding; its values are found depending on the circumferential

Table 6.8 Coefficients K_{HV} of the internal dynamics of loading in contact stress analysis. The values for the spurs are given in the numerator and the values for the helical wheels are given in the denominator

Accuracy degree according to GOST 1643-81	Hardness on the teeth surface	Values of K_{HV} in v (m/s)				
		1	3	5	8	10
6	> 350 HB	1.02 / 1.01	1.06 / 1.03	1.10 / 1.04	1.16 / 1.06	1.20 / 1.08
	≤ 350 HB	1.03 / 1.01	1.09 / 1.03	1.16 / 1.06	1.25 / 1.09	1.32 / 1.13
7	> 350 HB	1.02 / 1.01	1.06 / 1.03	1.12 / 1.05	1.19 / 1.08	1.25 / 1.10
	≤ 350 HB	1.04 / 1.02	1.12 / 1.06	1.20 / 1.08	1.32 / 1.13	1.40 / 1.16
8	> 350 HB	1.03 / 1.01	1.09 / 1.03	1.15 / 1.06	1.24 / 1.09	1.30 / 1.12
	≤ 350 HB	1.05 / 1.02	1.15 / 1.06	1.24 / 1.10	1.38 / 1.15	1.48 / 1.19
9	> 350 HB	1.03 / 1.01	1.09 / 1.03	1.17 / 1.07	1.28 / 1.11	1.35 / 1.14
	≤ 350 HB	1.06 / 1.02	1.12 / 1.06	1.28 / 1.11	1.45 / 1.18	1.56 / 1.22

Table 6.9 Imbalance factors $K^{\circ}_{H\beta}$ along the contact lines

ψ_{bd}	Hardness on the surrface of the wheel teeth	The values $K^{\circ}_{H\beta}$ for the gearing layout according to Fig. 6.37						
		1	2	3	4	5	6	7
0.4	≤ 350 HB	1.17	1.12	1.05	1.03	1.02	1.02	1.01
	> 350 HB	1.43	1.24	1.11	1.08	1.05	1.02	1.01
0.6	≤ 350 HB	1.27	1.18	1.08	1.05	1.04	1.03	1.02
	> 350 HB	–	1.43	1.20	1.13	1.08	1.05	1.02
0.8	≤ 350 HB	1.45	1.27	1.12	1.08	1.05	1.03	1.02
	> 350 HB	–	–	1.28	1.20	1.13	1.07	1.04
1.0	≤ 350 HB	–	–	1.15	1.10	1.07	1.04	1.02
	> 350 HB	–	–	1.38	1.27	1.18	1.11	1.06
1.2	≤ 350 HB	–	–	1.18	1.13	1.08	1.06	1.03
	> 350 HB	–	–	1.48	1.34	1.25	1.15	1.08
1.4	≤ 350 HB	–	–	1.23	1.17	1.12	1.08	1.04
	> 350 HB	–	–	–	1.42	1.31	1.20	1.12
1.6	≤ 350 HB	–	–	1.28	1.20	1.15	1.11	1.06
	> 350 HB	–	–	–	–	–	1.26	1.16

Table 6.10 Run-in coefficients K_{Hw} of gearings

Hardness on the teeth surface	Values K_{Hw} in v (m/s)					
	1	3	5	8	10	15
200 HB	0.19	0.20	0.22	0.27	0.32	0.54
250 HB	0.26	0.28	0.32	0.39	0.45	0.67
300 HB	0.35	0.37	0.41	0.50	0.58	0.87
350 HB	0.45	0.46	0.53	0.64	0.73	1.00
43 HRC	0.53	0.57	0.63	0.78	0.91	1.00
47 HRC	0.63	0.70	0.78	0.98	1.00	1.00
51 HRC	0.71	0.90	1.00	1.00	1.00	1.00
60 HRC	0.80	0.90	1.00	1.00	1.00	1.00

velocity of the gear wheel with the lower hardness (Table 6.10).

The value of the factor $K^0_{H\alpha}$ is determined depending on the accuracy degree ($n_{ac} = 5, 6, 7, 8, 9$) according to *smoothness standards*:

- For spur gears $K^0_{H\alpha} = 1$
- For helical gearings $K^0_{H\alpha} = 1 + A(n_{ac} - 5)$

where $A = 0.12$ for gear wheels with hardness H_1 and $H_2 > 350$ HB, and $A = 0.06$ for H_1 and $H_2 \leq 350$ HB or $H_1 > 350$ HB, and $H_2 \leq 350$ HB.

In the case of large-scale manufacture of reduction gears the computed value a_w is approximated to the nearest standard value: 50, 63, 71, 80, 90, 100, 112, 125, 140, 160, 180, 200, 224, 250, 260, 280, 300, 320, 340, 360, 380, and 400 mm.

5. The preliminary basic wheel dimensions are:
   - Pitch diameter $d_2 = 2a_w u/(u \pm 1)$
   - Width $b_2 = \psi_{ba} a_w$
6. The gear module. The maximum allowed module m_{max} (mm) is determined from the condition of no teeth undercutting at the root

$$m_{max} \approx 2a_w/[17(u \pm 1)] .$$

The minimum value of the module m_{min} (mm) is determined from the strength condition

$$m_{min} = \frac{K_m K_F T_1 (u \pm 1)}{a_w b_2 [\sigma]_F} ,$$

where $K_m = 3.4 \times 10^3$ for spur gears and $K_m = 2.8 \times 10^3$ for helical gears, and instead of $[\sigma]_F$ the lowest of the values of $[\sigma]_{F2}$ and $[\sigma]_{F1}$ is substituted.

The load factor in the bending stress analysis is

$$K_F = K_{FV} K_{F\beta} K_{F\alpha} .$$

The factor K_{FV} takes the internal loading dynamics into account. The values of K_{FV} are taken from Table 6.12 and depend on the accuracy degree according to the smoothness standards, the circumferential velocity, and the hardness of the working surfaces. The coefficient $K_{F\beta}$, which considers the unevenness of the stress distribution at the teeth root along the face width, is evaluated in accordance with

$$K_{F\beta} = 0.18 + 0.82 K^0_{H\beta} .$$

The coefficient $K_{F\alpha}$, which considers the influence of manufacturing errors in the pinion and the wheel on the load distribution between the teeth, is determined in the same way as in contact strength analysis: $K_{F\alpha} = K^0_{H\alpha}$.

From the given range (m_{min}–m_{max}) of the modules the lowest value m is taken, adjusting it with stan-

Table 6.12 Coefficients K_{FV} of the internal dynamics of loading in bending stress analysis. The values for the spurs are given in the numerator, and the values for the helical wheels are in the denominator

Accuracy degree according to GOST 1643-81	Hardness on the surface of the wheel teeth	The values K_{FV} in v (m/s)				
		1	3	5	8	10
6	> 350 HB	1.02 1.01	1.06 1.03	1.10 1.06	1.16 1.06	1.20 1.08
	≤ 350 HB	1.06 1.03	1.18 1.09	1.32 1.13	1.50 1.20	1.64 1.26
7	> 350 HB	1.02 1.01	1.06 1.03	1.12 1.05	1.19 1.08	1.25 1.10
	≤ 350 HB	1.08 1.03	1.24 1.09	1.40 1.16	1.64 1.25	1.80 1.32
8	> 350 HB	1.03 1,01	1.09 1.03	1.15 1.06	1.24 1.09	1.30 1.12
	≤ 350 HB	1.10 1.04	1.30 1.12	1.48 1.19	1.77 1.30	1.96 1.38
9	> 350 HB	1.03 1.01	1.09 1.03	1.17 1.07	1.28 1.11	1.35 1.14
	≤ 350 HB	1.11 1.04	1.33 1.12	1.56 1.22	1.90 1.36	– 1.45

dard ones (series 1 is preferable to series 2):

Table 6.11 Standard modules values

Series 1 (mm)	Series 2 (mm)
1.0	1.125
1.25	1.375
1.5	1.75
2.0	2.25
2.5	2.75
3.0	3.5
4.0	4.5
5.0	5.5
6.0	7.0
8.0	9.0
10.0	–

Values of the modules $m < 1\,\text{mm}$ with hardness $H \leq 350\,\text{HB}$ and $m < 1.5\,\text{mm}$ with hardness $H \geq 40\,\text{HRC}$ are undesirable for power trains.

7. The total teeth number and helix angle. The minimum tilt teeth angle of helical wheels amounts to

$$\beta_{\min} = \arcsin(3.45m/b_2)\,,$$

for herring-bone gears $\beta_{\min} = 25°$. The total number of the teeth becomes

$$z_\Sigma = 2a_w \cos\beta_{\min}/m\,.$$

The obtained value of z_Σ is rounded down to a whole number and the real value of the tilt tooth angle β is given by

$$\beta = \arccos[z_\Sigma m/(2a_w)]\,.$$

For helical gears $\beta = 8{-}20°$ whereas for herring-bone gears $\beta = 25{-}40°$.

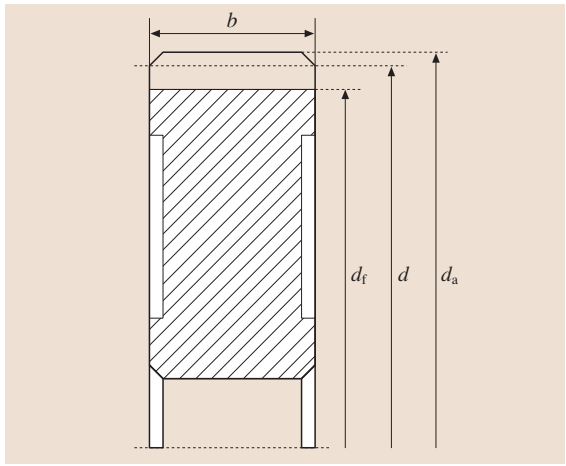


Fig. 6.38 Main diameters of the wheel

8. The teeth number of the wheel and pinion. The teeth number of the pinion is yielded as

$$z_1 = z_\Sigma/(u \pm 1) \geq z_{1\,\min}\,.$$

The value z_1 is rounded up to a whole number. For spurs $z_{1\,\min} = 17$, whereas for helical and herring-bone gears $z_{1\,\min} = 17\cos^3\beta$.
For $z_1 < 17$ the gearing is made with shifting to avoid teeth undercutting and to increase their breaking strength. The coefficient of displacement is

$$x_1 = (17 - z_1)/17 \leq 0.6\,.$$

For wheels with external toothing $x_2 = -x_1$, whereas for wheels with internal toothing $x_2 = x_1$. The teeth number of wheels with external toothing becomes $z_2 = z_\Sigma - z_1$, whereas the teeth number of wheels with internal toothing becomes $z_2 = z_\Sigma + z_1$.

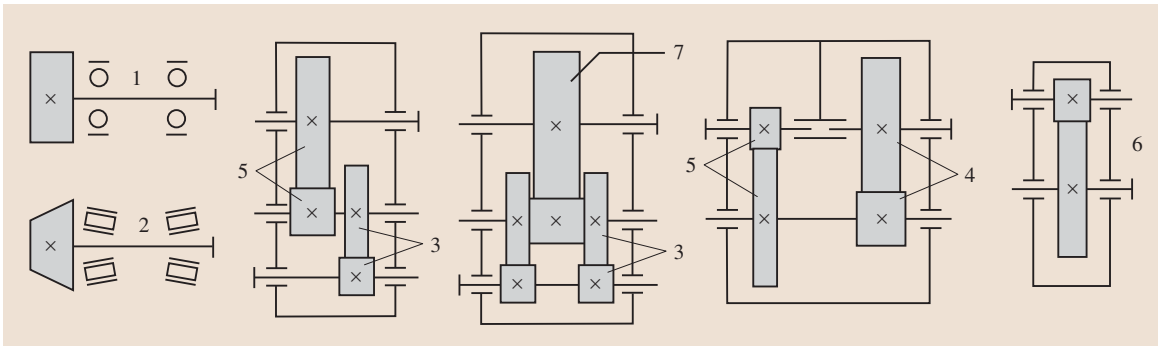


Fig. 6.37 Principal diagrams of reduction gears

9. The actual gear ratio. The actual gear ratio is $u_r = z_2/z_1$. The actual value of the gear ratio should not differ from the nominal ones by more than 3% for single-reduction units, by 4% for double-reduction units, and by 5% for multistage reduction units.
10. Wheel diameters (Fig. 6.38)
Pitch diameters d:

of the pinion $d_1 = z_1 m/\cos\beta$,
of the wheel with external toothing $d_2 = 2a_w - d_1$,
of the wheel with internal toothing $d_2 = 2a_w + d_1$.

Table 6.13 Formulas for calculation of the pitch diameters d

Wheel type	Calculating formulas
Pinion	$d_1 = z_1 m/\cos\beta$
Externally toothing wheel	$d_2 = 2a_w - d_1$
Internal toothing wheel	$d_2 = 2a_w + d_1$

The diameters d_a and d_f of the addendum circles and dedendum circles of the wheel teeth are:

– With external toothing

$$d_{a1} = d_1 + 2(1 + x_1 - y)m\ ;$$
$$d_{f1} = d_1 - 2(1.25 - x_1)m\ ;$$
$$d_{a2} = d_2 + 2(1 + x_2 - y)m\ ;$$
$$d_{f2} = d_2 - 2(1.25 - x_2)m\ ;$$

– With internal toothing

$$d_{a1} = d_1 + 2(1 + x_1)m\ ;$$
$$d_{f1} = d_1 - 2(1.25 - x_1)m\ ;$$
$$d_{a2} = d_2 - 2(1 - x_2 - 0.2)m\ ;$$
$$d_{f2} = d_2 + 2(1.25 + x_2)m\ ,$$

where x_1 and x_2 are coefficients of displacement of the pinion and the wheel, respectively, $y = -(a_w - a)/m$ is a coefficient of effective addendum modification, and a (the pitch center-to-center distance) is $a = 0.5\,m\,(z_2 \pm z_1)$.

11. Blank part dimensions
To obtain acceptable mechanical characteristics of the wheel material during heat treatment, the calculation requires that uncut dimensions of the wheels D_{blank}, C_{blank}, S_{blank} do not exceed the maximum allowable values D_{max}, S_{max} (Table 6.1)

$$D_{blank} \le D_{max}\ ;$$
$$C_{blank} \le S_{max} \quad \text{or}$$
$$S_{blank} \le S_{max}\ .$$

The values D_{blank}, C_{blank}, and S_{blank} (mm) are determined from the following formulas: for spur pinions (Fig. 6.39a) $D_{blank} = d_a + 6$ mm; for bevel pinion (Fig. 6.39b) $D_{blank} = d_{ae} + 6$ mm; for wheels with recesses (Fig. 6.39c) $C_{blank} = 0.5 b_2$ and $S_{blank} = 8$ m; and for wheels without recesses (Fig. 6.38) $S_{blank} = b_2 + 4$ mm.

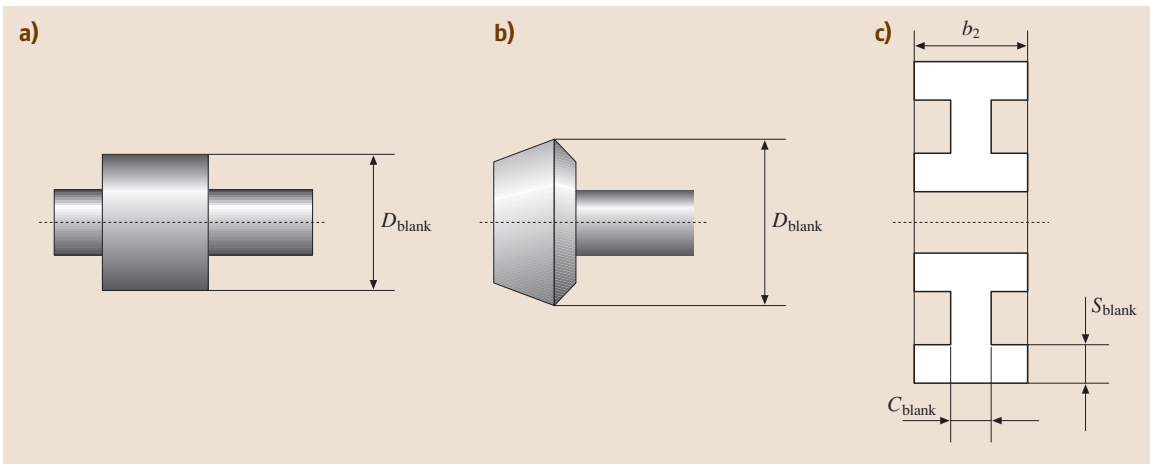


Fig. 6.39a–c Dimensions of the workpieces: (a) for a spur pinion, (b) for a bevel pinion, and (c) for a wheel with necks

Nonfulfillment of these inequalities means that the material of the details and the heat treatment method must be changed.

12. Contact stress testing of wheel teeth
The calculated value of the contact stress is determined from (6.6). If the calculated stress σ_H is lower than the allowable stress $[\sigma]_H$ by 15–20%, or if σ_H is higher than $[\sigma]_H$ by 5% or less, then the earlier accepted parameters of the gearing are taken as final. Otherwise recalculation is required.
13. Forces in toothing (Fig. 6.29) are given by

$$\textit{peripheral } F_t = 2\times 10^3 T_1/d_1 \ ;$$
$$\textit{radial } F_r = F_t \tan\alpha/\cos\beta \ ,$$
$$(\text{for the standard angle } \alpha = 20^\circ \ \tan\alpha = 0.364) \ ;$$
$$\textit{axial } F_a = F_t \tan\beta \ .$$

14. Bending stress testing of wheel teeth
The calculated bending stress in the wheel teeth is

$$\sigma_{F2} = \frac{K_F F_t}{b_2 m} Y_{FS2} Y_\beta Y_\varepsilon \le [\sigma]_{F2} \ ,$$

and in the pinion teeth

$$\sigma_{F1} = \sigma_{F2} Y_{FS1}/Y_{FS2} \le [\sigma]_{F1} \ .$$

The value of the coefficient Y_{FS} takes into account the form of the tooth and stress concentration depending on the reduced number $z_v = z/\cos^3\beta$ of the teeth. The coefficient x of the displacement for external toothing is taken according to Table 6.14.

For internal toothing it is as follows:

Table 6.15 Coefficient Y_{FS} for internally toothing

z	Y_{FS}
40	4.02
50	3.88
63	3.80
71	3.75

The value of the coefficient Y_β, which considers the tilt angle of the tooth in *helical* gearings, is determined from (β in degrees)

$$Y_\beta = 1 - \varepsilon_\beta \beta/120 \ ,$$

under the condition that $Y_\beta \ge 0.7$. Y_ε is a coefficient that takes teeth overlap into account.
For spur gears $Y_\beta = 1$; $Y_\varepsilon = 1$ in the case of accuracy degrees 8 or 9; $Y_\varepsilon = 0.8$ applies for accuracy degrees 5–7. For helical gearings one has $Y_\varepsilon = 0.65$.

15. Checking strength calculation of teeth under peak load
The aim of this calculation is prevention of residual strains or brittle fracture of the surface layer or the teeth under the action of the maximum torque T_{max}.
The action of peak loads is evaluated by means of the overload factor $K_{load} = T_{max}/T$, where T is the nominal torque.
For prevention of residual strains or brittle fracture of the surface layer the contact stress $\sigma_{H\,max}$ must not exceed the allowable stress $[\sigma]_{H\,max}$

$$\sigma_{H\,max} = \sigma_H \sqrt{K_{load}} \le [\sigma]_{H\,max} \ ,$$

where σ_H is the contact stress under the action of the nominal torque T.

Table 6.14 Coefficients Y_{FS} of the tooth form and stress concentration

z or z_v	Values of Y_{FS} for the coefficient of displacement x of the tools						
	−0.6	−0.4	−0.2	0	+0.2	+0.4	+0.6
12	–	–	–	–	–	3.67	–
14	–	–	–	–	4.00	3.62	3.30
17	–	–	–	4.30	3.89	3.58	3.32
20	–	–	–	4.08	3.78	3.56	3.34
25	–	–	4.22	3.91	3.70	3.52	3.37
30	–	4.38	4.02	3.80	3.64	3.51	3.40
40	4.37	4.06	3.86	3.70	3.60	3.51	3.42
60	3.98	3.80	3.70	3.62	3.57	3.52	3.46
80	3.80	3.71	3.63	3.60	3.57	3.53	3.49
100	3.71	3.66	3.62	3.59	3.58	3.53	3.51
200	3.62	3.61	3.61	3.59	3.59	3.59	3.56

The allowable stress $[\sigma]_{\mathrm{H\,max}}$ for refining or through quenching is taken to be $[\sigma]_{\mathrm{H\,max}} = 2.8\sigma_y$, for cementing or induction hardening current it is $[\sigma]_{\mathrm{H\,max}} = 44\,\mathrm{HRC_m}$, and for nitriding it is $[\sigma]_{\mathrm{H\,max}} \approx 35\,\mathrm{HRC_m} \leq 2000\,\mathrm{N/mm^2}$.
For the prevention of residual strains and brittle fracture of the teeth the bending stress $\sigma_{\mathrm{F\,max}}$ under the action of the maximum torque must not exceed the allowable value of $[\sigma]_{\mathrm{F\,max}}$, i.e.,

$$\sigma_{\mathrm{F\,max}} = \sigma_{\mathrm{F}} K_{\mathrm{load}} \leq [\sigma]_{\mathrm{F\,max}} ,$$

where σ_{F} is the bending stress determined from the fatigue strength calculations. Testing is carried out separately for the pinion and the wheel. The allowable limit stress of the bending is computed depending on the heat treatment method applied and the potential application frequency of the peak load

$$[\sigma]_{\mathrm{F\,max}} = \sigma_{\mathrm{F\,lim}} Y_{\mathrm{N\,max}} k_{\mathrm{st}} / S_{\mathrm{st}} ,$$

where $\sigma_{\mathrm{F\,lim}}$ is the fatigue bending point. $Y_{\mathrm{N\,max}} = 4$ for steel with volume heat treatment (normalization, refining, and volume quenching). $Y_{\mathrm{N\,max}} = 2.5$ for steel with surface treatment (quenching with heating by means of high-frequency currents, cementing, nitrocementing, and nitriding). k_{st} is a coefficient that captures the influence of the frequency of the peak load application; in the case of single overloads ($\leq 10^3$) $k_{\mathrm{st}} = 1.2$–1.3 are high values for volume heat treatment; under repeated ($> 10^3$) overload action $k_{\mathrm{st}} = 1$. S_{st} is a load factor; usually $S_{\mathrm{st}} = 1.75$.

6.4 Bevel Gearings

6.4.1 Basic Considerations

Bevel gearings are used for the transmission of mechanical energy between shafts with intersecting axes. *Right-angle gears* (with angle $\Sigma = 90°$) are the most common (Fig. 6.40). As has been noted bevel wheels can have *straight* or *circular* teeth.

The tooth lines in bevel wheels with circular teeth are circular arcs. Gearings with straight teeth have initial *linear contact*, and those with circular teeth have *point contact* in the toothing. The tilt angle β_n of the tooth line is determined in the average section along the face width. For gearings with straight teeth $\beta_n = 0$; for those with circular teeth $\beta_n = 35°$. The tilt of the tooth line increases the operating smoothness, and the contact and bending strengths, but also increases the load on the bearings and shafts. Bevel wheels with circular teeth as compared with spurs have a higher load-carrying ability, and operate smoothly and with less noise. When, the total contact ratio $\varepsilon_\gamma = \sqrt{\varepsilon_\alpha^2 + \varepsilon_\beta^2}$, is more than two, not less than two tooth pairs of teeth are constantly operating.

In order to increase the wear and sliding strength of teeth, specific slips in the boundary points of the toothing are adjusted by means of shifting of the initial contour. The pinion is made with positive shifting, and the wheel with negative shifting of the same absolute value.

The analog of the pitch cylinder of cylindrical gearings for bevel gearings are pitch cones that are coincident with the initial cones. When the wheels rotate, pitch cones move against one another without slip. Bevel gearings should be adjusted to achieve coincidence of the pitch cone points of the wheels. The angle Σ between the axes of the gearings equals the angular sum of the pitch cones $\Sigma = \delta_1 + \delta_2$ (Fig. 6.40).

The advantage of bevel gearings is their ability to transmit mechanical power between shafts with inter-

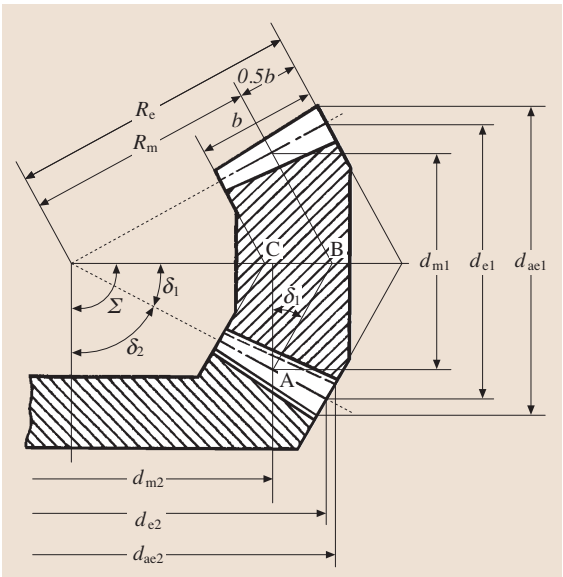


Fig. 6.40 Geometry of a bevel wheel

secting axes. Their disadvantages are the necessity for adjustment of the gearing (the pitch cone points must coincide), as well as their lower load-carrying capacity and greater manufacture complexity as compared with cylindrical gearings.

The external and internal ends of bevel wheels form external and internal *extra cones*, of which the generating lines are perpendicular to the generating line of the pitch cone. The mean extra cone is positioned an equal distance from the external and internal extra cones. The face width b of the gear wheel is limited by two extra cones – external and internal. The length of the generating line from its point to its external end is called the *external cone distance* R_e, and the length to the middle of the face width is called the *average cone distance* R_m (Fig. 6.40). The intersections of the pitch cone with extra cones determine the diameters of the pitch circles of the bevel wheel, and distinguish the external d_e, internal d_i, and mean d_m pitch diameters.

According to Fig. 6.40 the gear ratio becomes

$$u = d_{e2}/d_{e1} = d_{m2}/d_{m1} = \tan\delta_2 = 1/\tan\delta_1 = z_2/z_1 ,$$

where d_{e1}, d_{e2}, d_{m1}, d_{m2}, and δ_1, δ_2 are, respectively, the external and mean pitch diameters, and pitch cone angles, of the pinion (subscript "1") and the wheel (subscript "2"). For bevel spur gears $u = 2–3$ is recommended, whereas for wheels where the circular teeth u has a value as high as 6.3.

6.4.2 The Axial Tooth Form

The teeth of bevel wheels depend on dimensional changes of their normal sections along their length and are made of three axial forms (Fig. 6.41)

Axial Form I – Normally Reducing Teeth (Fig. 6.41a)

In this form, the pitch cone and dedendum cone points coincide, and the root is proportionate to the cone radius. This is used for straight teeth as well as in restricted cases for circular teeth with $m \geq 2$ mm and $\sqrt{z_1^2 + z_2^2} = 20–100$.

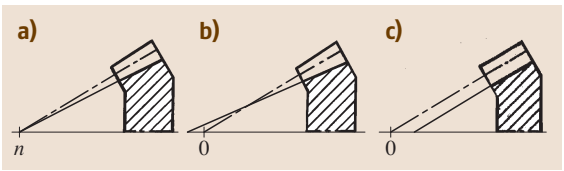


Fig. 6.41a–c Axial tooth form of the bevel wheel: **(a)** normally lowered (I), **(b)** normally narrowed (II), and **(c)** uniformly high (III)

Axial Form II – Normally Convergent Teeth (Fig. 6.41b)

In this form, the dedendum cone point is positioned so that the bottom land width is constant and the tooth thickness along the pitch cone is proportionate to the cone radius. This form provides the optimal bending strength in all sections, allows both surfaces of the gear wheels to be treated with one tool at once, and thus increases productivity in the gear cutting. This form is the basic one used for wheels with circular teeth and is used in mass production.

Axial Form III – Teeth of Equal Depth (Fig. 6.41c)

In this form, the generating lines of the pitch cone, the dedendum cone, and the addendum cone are parallel. The teeth depth is constant along the whole length. This is used for nonorthogonal gearings with axis angle $\Sigma < 40°$ and circular teeth with $\sqrt{z_1^2 + z_2^2} \geq 60$.

6.4.3 Basic Geometric Proportions

In bevel wheels with axial forms I and II the tooth depth and therefore the toothing module increase from the internal cone to the external extra cone (Figs. 6.40 and 6.41). For measurement convenience it is customary to define the dimensions of bevel wheels according to the external tooth end. The maximum teeth module – the exterior peripheral module m_{te} – is received on the external wheel end. The basic geometric proportions for bevel gearings are listed below (Fig. 6.40).

The outer pitch diameters of the pinion and the wheel are

$$d_{e1} = m_{te} z_1; \quad d_{e2} = m_{te} z_2 .$$

The outer cone radius is

$$R_e = \sqrt{(0.5 d_{e1})^2 + (0.5 d_{e2})^2} = 0.5 d_{e1}\sqrt{1+u^2} .$$

The face width is $b = K_{be} R_e$. For most bevel gearings the face width coefficient is $K_{be} = 0.285$. Then

$$b = 0.285 \times 0.5 d_{e1}\sqrt{1+u^2} = 0.143 d_{e1}\sqrt{1+u^2} .$$

The average cone radius then becomes

$$R_m = R_e - 0.5b = R_e - 0.5 \times 0.285 R_e = 0.857 R_e .$$

From the similarity condition (Fig. 6.40) it follows that $d_{e1}/R_e = d_{m1}/R_m$. Then the mean pitch diameter of the

pinion becomes

$$d_{m1} = d_{e1} R_m / R_e = 0.857 d_{e1} .$$

The peripheral module in the mean section $m_{tm} = 0.857 m_{te}$. The normal module in the mean section for the circular tooth ($\beta_n = 35°$) is

$$m_n = m_{tm} \cos \beta_n \approx 0.702 m_{te} .$$

The pitch cone angles are

$$\tan \delta_1 = z_1 / z_2 = 1/u; \quad \delta_2 = 90° - \delta_1 .$$

The external peripheral module m_{te} is taken as a rated module for bevel wheels with straight teeth; the mean normal module m_n is taken for bevel wheels with circular teeth in the middle of the gear ring.

The teeth of the bevel wheels with modules that change over a certain continuous range can be cut with the same gear-shaping cutter head. Thus, rogue values of the module can be used.

6.4.4 Equivalent Cylindrical Wheels

For the spur, the gear teeth profiles of the bevel wheel on the middle extra cone (Fig. 6.42) are similar to the teeth

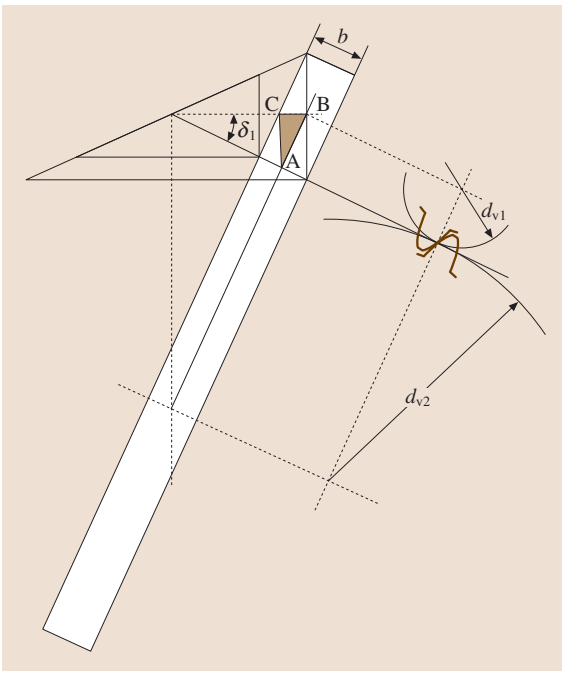


Fig. 6.42 Reduction of the bevel wheel to the equivalent spur

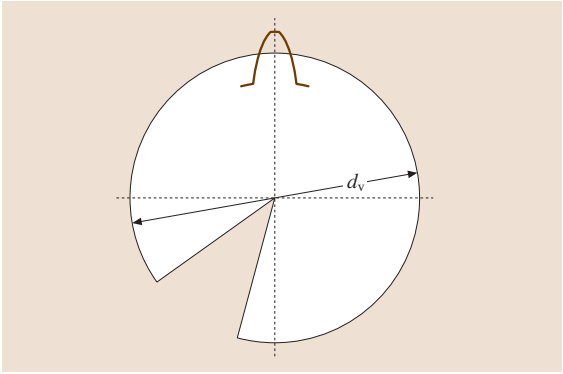


Fig. 6.43 Development of the middle additional cone on the plane

profiles of a spur gear with pitch diameter d_v. By adding an extra middle cone on the plane (Fig. 6.43) up to the whole cycle, we obtain an equivalent cylindrical wheel with teeth number z_v and pitch diameter $d_v = m_n z_v$. After consideration of the triangle ABC (Fig. 6.42) we can obtain a relation between the pitch diameters d_v and d_m

$$d_v = d_m / \cos \delta = m_n z / \cos \delta .$$

From the equality $m_n z_v = m_n z / \cos \delta$ a relation for the determination of the *equivalent teeth number* follows as

$$z_v = z / \cos \delta ,$$

i. e., the actual straight bevel gear with teeth number z in the strength analysis can be replaced with a cylindrical one with teeth number z_v.

For gearings with circular teeth, the teeth profiles of the bevel wheel in the normal section are similar to those of the equivalent spur gear. The equivalent teeth number z_{vn} is obtained by means of double reduction: of the bevel wheel to the cylindrical one, and of the circular tooth to the straight tooth, as follows

$$z_{vn} = z / \left(\cos \delta \cos^3 \beta_n \right) .$$

6.4.5 Toothing Forces

In the bevel gearing, the site of the force application F_n, which acts normal on the tooth surface, is considered to be the section in the middle of the face width.

For calculation for the shafts and bearings it is convenient for the force F_n to be represented in the form of its constituents F_t, F_r, and F_a.

The peripheral force F_t (N) on the pinion becomes

$$F_t = 2 \times 10^3 T_1 / d_{m1} ,$$

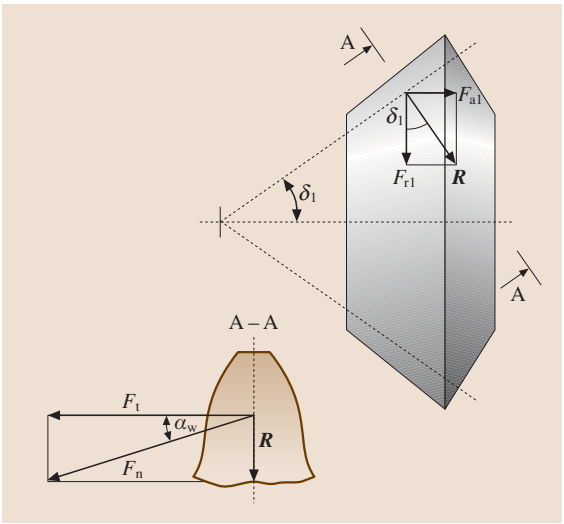


Fig. 6.44 Forces loading the bevel pinion

where T_1 is the torque (N m) and d_{m1} is the mean pitch diameter (in mm).

To determine the constituents for spur gears (Fig. 6.44), let us write the intermediate expression (where $\alpha_w = 20°$, the angle of action)

$$R = F_t \tan \alpha_w \, .$$

The radial force on the pinion is then

$$F_{r1} = R \cos \delta_1 = F_t \tan \alpha_w \cos \delta_1 \, .$$

The axial force on the pinion becomes

$$F_{a1} = R \sin \delta_1 = F_t \tan \alpha_w \sin \delta_1 \, .$$

The corresponding forces on the wheel equal (Fig. 6.45) are then $F_{r2} = F_{a1}$ and $F_{a2} = F_{r1}$. In gearings with circular teeth it is necessary to ensure that the axial force F_{a1} on the drive pinion is directed towards the base of the pitch cone. This, in order to avoid teeth seizing, caused by considerable gaps in the bearings. The rotation of the drive pinion (as seen from the direction of the vertex of the pitch cone) and the direction of the teeth helix angle must coincide. In Fig. 6.46 the pinion is rotating counterclockwise, i. e., to the left, and the tooth of the pinion is left tooth.

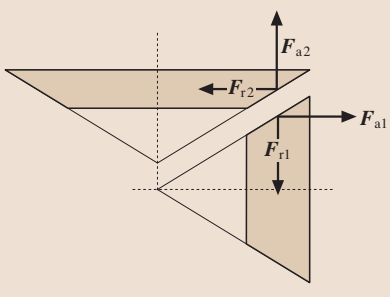


Fig. 6.45 Forces in the toothing of a bevel gearing

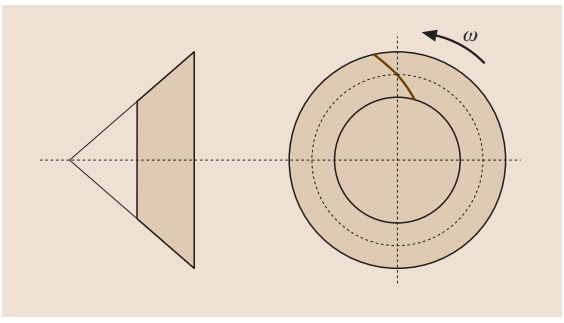


Fig. 6.46 Recommended direction of tooth dip for the drive pinion

If this condition is met, in gearings with circular teeth the radial force of the pinion is

$$F_{r1} = F_t(\tan \alpha_w \cos \delta_1 - \sin \beta_n \sin \delta_1)/\cos \beta_n \, .$$

The axial force on the pinion is

$$F_{a1} = F_t(\tan \alpha_w \sin \delta_1 + \sin \beta_n \cos \delta_1)/\cos \beta_n \, .$$

The signs in the formulas will be the same in the case of clockwise rotation of the drive pinion with the tooth on the right. The forces on the wheel equal, respectively, $F_{r2} = F_{a1}$ and $F_{a2} = F_{r1}$.

6.4.6 Contact Strength Analysis of Bevel Gearings

Strength analysis of bevel gearings is based on the assumption that the load-carrying capacity of the teeth of bevel wheels is the same as that of the equivalent cylindrical wheels with the same tooth length b and a profile corresponding to the middle extra cone (the mean tooth section).

Checking Analysis

Equation (6.6), expressed using the parameters of the equivalent cylindrical spur gear according to the mean extra cone (Fig. 6.40), is

$$\sigma_H = Z_E Z_H Z_\varepsilon \sqrt{\frac{K_H F_t}{b d_{v1}} \frac{(u_v + 1)}{u_v} \frac{1}{\vartheta_H}} \, , \qquad (6.8)$$

where ϑ_H is the ratio taking into consideration the influence of the gearing of the bevel wheel type (i.e., straight or circular teeth) on the load-carrying capacity.

The gear ratio of the equivalent cylindrical gearing is then

$$u_v = \frac{d_{v2}}{d_{v1}} = \frac{d_{m2}}{\cos\delta_2}\frac{\cos\delta_1}{d_{m1}} = \frac{u\cos\delta_1}{\cos\delta_2} .$$

Considering that $\cos\delta_1 = \sin\delta_2$ (Fig. 6.40) and $\tan\delta_2 = u$, we have

$$u_v = u\sin\delta_2/\cos\delta_2 = u^2 .$$

The diameter of the equivalent spur pinion is $d_{v1} = d_{m1}/\cos\delta_1$. Substituting the cosine function for the tangent function we obtain

$$\cos\delta_1 = 1/\sqrt{1+\tan^2\delta_1} .$$

Bearing in mind that $\tan\delta_1 = 1/u$ and $d_{m1} = 0.857d_{e1}$ we can write

$$d_{v1} = d_{m1}/\cos\delta_1 = d_{m1}\sqrt{1+\tan^2\delta_1}$$
$$= d_{m1}\sqrt{\left(u^2+1\right)/u^2} = 0.857d_{e1}\sqrt{1+u^2}/u .$$

Substituting the values u_v and d_{v1} into (6.8) and replacing $F_t = 2\times10^3 T_1/(0.857d_{e1})$, $b = 0.143d_{e1}\sqrt{1+u^2}$, subject to the strength condition $\sigma_H \le [\sigma]_H$ we obtain the formula for checking analysis of steel bevel gearings

$$\sigma_H = 6.7\times10^4\sqrt{\frac{K_H T_1}{d_{e1}^3 u\vartheta_H}} \le [\sigma]_H , \qquad (6.9)$$

where T_1 is in N m, d_{e1} is in mm, and σ_H and $[\sigma]_H$ are in N/mm^2. The load factor K_H for bevel gearings is $K_H = K_A K_{H\beta} K_{HV}$. The values of the ratio K_A are set in the same way as for cylindrical gearings. The factor $K_{H\beta}$ takes the unevenness of load distribution along the contact lines into account. K_{HV} considers internal dynamic load.

Checking Analysis

Having solved (6.9) relative to d_{e1} we obtain the checking analysis formula for steel bevel gearings as

$$d_{e1} = 1650\sqrt[3]{\frac{K_H T_1}{u[\sigma]_H^2\vartheta_H}} ,$$

where d_{e1} is an outer pitch diameter of the pinion (mm), T_1 is in N m, and $[\sigma]_H$ is in N/mm^2.

6.4.7 Calculation of the Bending Strength of Bevel Gearing Teeth

Similarly as for straight cylindrical gearings, the bending strength condition is checked for the teeth of the pinion and the wheel

$$\sigma_{F1} = \frac{K_F F_t Y_{FS1}}{b m_n \vartheta_F} \le [\sigma]_{F1} ;$$
$$\sigma_{F2} = \frac{Y_{FS2}}{Y_{FS1}}\sigma_{F1} \le [\sigma]_{F2} ,$$

where K_F is the load factor, m_n is the normal module in the mean section of the bevel wheel, Y_{FS} is the ratio of the tooth form and stress concentration of the equivalent wheel [Y_{FS} is chosen according to z_v (z_{vn})], and ϑ_F is the ratio that takes into account the influence of the bevel-wheel gearing on the load-carrying capacity.

The load factor K_F for bevel gearings is

$$K_F = K_A K_{F\beta} K_{FV} .$$

The values of the coefficient K_A are obtained in the same way as for the cylindrical gearings. $K_{F\beta}$ is the ratio that considers the unevenness of the stress distribution at the teeth root along the face width, and K_{FV} is a coefficient for the internal dynamic load. The choice of the allowable stresses $[\sigma]_{F1}$ and $[\sigma]_{F2}$ was explained above.

6.4.8 Projection Calculation for Bevel Gearings

The following basic data are considered: T_1 (the torque on the pinion measured in N m), typical loading conditions, n_1 (the rotational frequency of the pinion measured in min^{-1}), u (the gear ratio), L_h (the operation time of the gearing, i.e., the lifetime, measured in hours), and the gearing layout, and gear wheel type. The choice of materials, heat treatment method, and the determination of the allowable stresses are given in Sect. 6.3.7.

Diameter of the Outer Pitch Circle of the Pinion

The tentative value of the outer pitch circle diameter of the pinion (mm) is

$$d'_{e1} = K\sqrt[3]{\frac{T_1}{u\vartheta_H}} ,$$

where T_1 is the torque on the pinion (N m) and u is the gear ratio. The factor K, which depends on the surface hardnesses H_1 and H_2 of the teeth of the pinion and the wheel, have the following values, respectively:

Table 6.16 Coefficient K for the bevel gearings

Hardness H		Factor K
$H_1 \le 350$ HB	$H_2 \le 350$ HB	30
$H_1 \ge 45$ HRC	$H_2 \le 350$ HB	25
$H_1 \ge 45$ HRC	$H_2 \ge 45$ HRC	22

The values of the factor ϑ_H are taken as:

- For straight bevel gearings $\vartheta_H = 0.85$
- For gearings with circular teeth, according to Table 6.17

The circumferential velocity v_m (m/s) at the mean pitch diameter is determined from (with $K_{be} = 0.285$)

$$v_m = \pi 0.857 d'_{e1} n_1 / (6 \times 10^4) .$$

The accuracy degree is specified depending on circumferential velocity. Straight bevel gears are used with circumferential velocities up to 5 m/s; the accuracy degree is not lower than 7. Bevel wheels with circular teeth with a circumferential velocity of up to 5 m/s are made with an accuracy degree of not lower than 8; with $v_m = 5$–10 m/s it is not lower than 7.

The previously determined diameter value of the outer pitch circle of the pinion is made more precise (in mm) using

$$d_{e1} = 1650 \sqrt[3]{\frac{K_{HV} K_{H\beta} T_1}{u \vartheta_H [\sigma]_H^2}} .$$

The factor K_{HV} of the internal dynamic load for *straight* bevel wheels is chosen from Table 6.8, conditionally taking their accuracy lower than actual the by one degree; i. e., instead of the actual accuracy degree 7 an accuracy degree of 8 is used for the factor K_{HV}. For bevel wheels with *circular teeth* the value of K_{HV} is taken from Table 6.8 as for helical wheels. The factor $K_{H\beta}$ considers the unevenness of the load distribution along the contact lines. In bevel gearings *the pinion is set as a cantilever*. In order to increase the rigidity of the bearing, the shafts are mounted *on tapered roller bearings*.

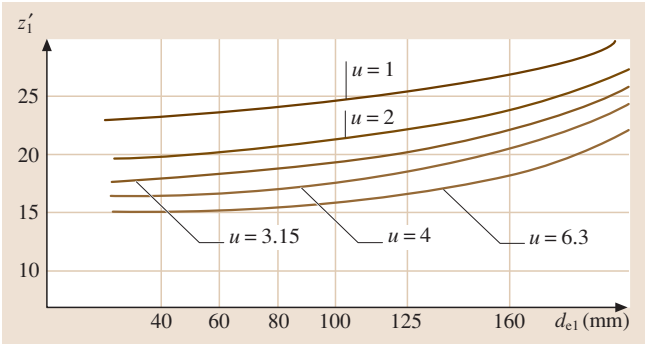


Fig. 6.47 Determination of the tooth number of a bevel pinion with straight teeth

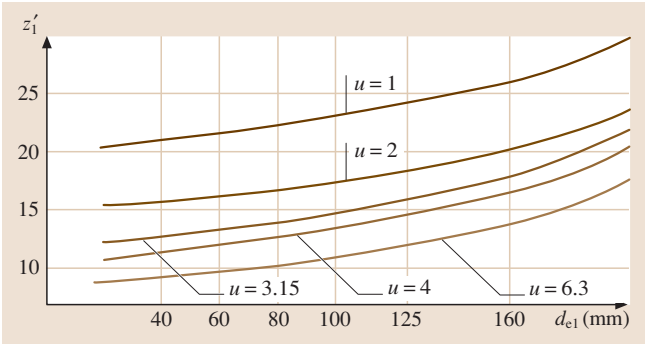


Fig. 6.48 Determination of the tooth number of a bevel pinion with circular teeth

For bevel wheels

with circular teeth $K_{H\beta} = \sqrt{K^0_{H\beta}}$,

on condition that $K_{H\beta} \ge 1.2$,

with straight teeth $K_{H\beta} = K^0_{H\beta}$.

The factor $K^0_{H\beta}$ is taken from Table 6.9 for cylindrical gearings, depending on the ratio $\psi_{bd} = b/d_{e1}$, the hardness of the gear wheels, and the gearing position relative to the bearings. As the face width and pinion diameter have not yet been determined, the value of the factor

Table 6.17 Coefficients ϑ_H and ϑ_F for bevel gearings with circular teeth

Hardness of the gear wheels		Values	
		ϑ_H	ϑ_F
$H_1 \le 350$ HB	$H_2 \le 350$ HB	$1.22 + 0.21u$	$0.94 + 0.08u$
$H_1 \ge 45$ HRC	$H_2 \le 350$ HB	$1.13 + 0.13u$	$0.85 + 0.04u$
$H_1 \ge 45$ HRC	$H_2 \ge 45$ HRC	$0.81 + 0.15u$	$0.65 + 0.11u$

ψ_{bd} is computed approximately from

$$\psi_{bd} = 0.166\sqrt{u^2+1}\,.$$

Cone Radius and Face Width

The angle of the pitch pinion cone is

$$\delta_1 = \arctan(1/u)\,.$$

The external cone distance is $R_e = d_{e1}/(2\sin\delta_1)$ and the face width is $b = 0.285R_e$.

The Gearing Module

The exterior end module of the gearing (m_e for bevel wheels with straight teeth, m_{te} for wheels with circular teeth) is

$$m_e\,(m_{te}) \geq \frac{14K_{FV}K_{F\beta}T_1}{d_{e1}b\vartheta_F[\sigma]_F}\,.$$

The value of the internal dynamic load factor K_{FV} for straight bevel wheels is chosen from Table 6.12, conditionally taking their accuracy as one degree rougher than the actual degree. For bevel wheels with circular teeth the value of K_{FV} is taken from Table 6.12, as for helical wheels.

The factor $K_{F\beta}$ takes the unevenness of the stress distribution at the teeth root along the face width into account. For bevel gearings with straight teeth one has $K_{F\beta} = K'_{F\beta}$; for wheels with circular teeth one uses

$$K_{F\beta} = \sqrt{K'_{F\beta}}\,,$$

on the condition that $K_{F\beta} \geq 1.15$, where $K'_{F\beta} = 0.18 + 0.82K^0_{H\beta}$. For spurs the coefficient ϑ_F is taken equal to 0.85 and for wheels with circular teeth it is taken from Table 6.17. Instead of $[\sigma]_F$ the lesser of $[\sigma]_{F1}$ and $[\sigma]_{F2}$ is substituted into the design formula. Rounding off of the computed module value to the standard value can be ignored.

Teeth Number of the Pinion and the Wheel

For the pinion with straight teeth one has $z_1 = d_{e1}/m_e$, whereas with circular teeth one has $z_1 = d_{e1}/m_{te}$.

The teeth number of the wheel is $z_2 = z_1u$. The given values are rounded to the whole number.

In practice, there is another method to determine the teeth number and wheel module. The tentative value of the teeth number of the pinion (z'_1) is chosen depending on its diameter d_{e1} and gear ratio u in accordance with one of the diagrams graphed for straight bevel wheels (Fig. 6.47) or wheels with circular teeth (Fig. 6.48), with the teeth hardness of the wheel and the pinion ≥ 45 HRC. z_1 is specified, taking into account the teeth hardness of the pinion and the wheel by:

Table 6.19 Correction of the pinion tooth number z_1

Hardness H		Teeth number z_1
$H_1 \leq 350$ HB	$H_2 \leq 350$ HB	$1.6z'_1$
$H_1 \geq 45$ HRC	$H_2 \leq 350$ HB	$1.3z'_1$
$H_1 \geq 45$ HRC	$H_2 \geq 45$ HRC	z'_1

The teeth number of the wheel is $z_2 = z_1u$. Calculated values of the teeth number of the pinion and the wheel are rounded to whole numbers. The exterior end module of the gearing is calculated (m_e for bevel wheels with straight teeth, m_{te} for wheels with circular teeth) by using

$$m_e(m_{te}) = d_{e1}/z_1\,.$$

Table 6.18 Coefficients of displacement x_{e1} for bevel pinions with straight teeth

z_1	x_{e1} for gear ratio u:							
	1.0	1.25	1.6	2.0	2.5	3.15	4.0	5.0
12	–	–	–	–	0.50	0.53	0.56	0.57
13	–	–	–	0.44	0.48	0.52	0.54	0.55
14	–	–	0.34	0.42	0.47	0.50	0.52	0.53
15	–	0.18	0.31	0.40	0.45	0.48	0.50	0.51
16	–	0.17	0.30	0.38	0.43	0.46	0.48	0.49
18	0.00	0.15	0.28	0.36	0.40	0.43	0.45	0.46
20	0.00	0.14	0.26	0.34	0.37	0.40	0.42	0.43
25	0.00	0.13	0.23	0.29	0.33	0.36	0.38	0.39
30	0.00	0.11	0.19	0.25	0.28	0.31	0.33	0.34
40	0.00	0.09	0.15	0.20	0.22	0.24	0.26	0.27

Table 6.20 Coefficients of displacement x_{n1} for bevel pinions with circular teeth

z_1	x_{n1} for the gear ratio u:							
	1.0	1.25	1.6	2.0	2.5	3.15	4.0	5.0
12	–	–	–	0.32	0.37	0.39	0.41	0.42
13	–	–	–	0.30	0.35	0.37	0.39	0.40
14	–	–	0.23	0.29	0.33	0.35	0.37	0.38
15	–	0.12	0.22	0.27	0.31	0.33	0.35	0.36
16	–	0.11	0.21	0.26	0.30	0.32	0.34	0.35
18	0.00	0.10	0.19	0.24	0.27	0.30	0.32	0.32
20	0.00	0.09	0.17	0.22	0.26	0.28	0.29	0.29
25	0.00	0.08	0.15	0.19	0.21	0.24	0.25	0.25
30	0.00	0.07	0.11	0.16	0.18	0.21	0.22	0.22
40	0.00	0.05	0.09	0.11	0.14	0.16	0.17	0.17

The Actual Gear Ratio $u_r = z_2/z_1$

The calculated value of u_r must not differ from the target value by more than 3% for bevel reduction gears, 4% for bevel-cylindrical double-reduction gears, and 5% for three-stage (or greater) bevel-cylindrical reduction gears.

Final Values of Wheel Dimensions

The pitch cone angles of the pinion and the wheel are

$$\delta_1 = \arctan(1/u_r); \quad \delta_2 = 90^\circ - \delta_1 .$$

The pitch diameters of the wheels are

with straight teeth $d_{e1} = m_e z_1$, $d_{e2} = m_e z_2$;

with circular teeth $d_{e1} = m_{te} z_1$, $d_{e2} = m_{te} z_2$.

The outer diameters of the wheels are

with straight teeth

$$d_{ae1} = d_{e1} + 2(1 + x_{e1}) m_e \cos\delta_1 ,$$

$$d_{ae2} = d_{e2} + 2(1 + x_{e2}) m_e \cos\delta_2 ;$$

with circular teeth

$$d_{ae1} = d_{e1} + 1.64(1 + x_{n1}) m_{te} \cos\delta_1 ,$$

$$d_{ae2} = d_{e2} + 1.64(1 + x_{n2}) m_{te} \cos\delta_2 .$$

The coefficients x_{e1} and x_{n1} for straight and helical *pinions* are taken from Tables 6.18 and 6.20. For gearings with z_1 and u that differ from those given in Tables 6.18 and 6.20, the values x_{e1} and x_{n1} are rounded up. The coefficient of tool displacement for the wheel is

$$x_{e2} = -x_{e1} ; \quad x_{n2} = -x_{n1} .$$

Uncut Wheel Dimensions

The dimensions of the billets are computed for the bevel pinion and the wheel (mm) (Fig. 6.39b) as

$$D_{blank} = d_{e1} + 2m_e(m_{te}) + 6\,\text{mm} ,$$

$$S_{blank} = 8m_e(m_{te}) .$$

The values of D_{blank} and S_{blank} determined from calculations are then compared with the limit dimensions D_{max} and S_{max} detailed in Table 6.1.

The conditions for suitability of the billets are

$$D_{blank} \le D_{max} ;$$

$$S_{blank} \le S_{max} .$$

Toothing Forces (Fig. 6.44)

The circumferential force on the mean diameter d_{m1} of the pinion is

$$F_t = 2 \times 10^3 T_1 / d_{m1} , \quad \text{where} \quad d_{m1} = 0.857 d_{e1} .$$

The axial force on the pinion is

with straight teeth $F_{a1} = F_t \tan\alpha_w \sin\delta_1$,

with circular teeth $F_{a1} = \gamma_a F_t$.

The radial force on the pinion is

with straight teeth $F_{r1} = F_t \tan\alpha_w \cos\delta_1$,

with circular teeth $F_{r1} = \gamma_r F_t$.

The axial force on the wheel is $F_{a2} = F_{r1}$, and the radial force on the wheel is $F_{r2} = F_{a1}$.

The coefficients γ_a and γ_r for the angle $\beta_n = 35°$ are determined from the formulas

$$\gamma_a = 0.44 \sin \delta_1 + 0.7 \cos \delta_1 ,$$

$$\gamma_r = 0.44 \cos \delta_1 - 0.7 \sin \delta_1 .$$

The calculated coefficients γ_a and γ_r are substituted into the formulas with their corresponding signs. Teeth seizing will not occur if the force F_{a1} is directed towards the base of the pitch cone of the drive pinion. Thus the rotating sense of the pinion (seen from the direction of the pitch cone point) and the dip direction of the teeth are chosen to be identical; e.g., for the drive pinion with a left tooth dip the sense of rotation is counterclockwise.

Contact Stress Analysis of Wheel Teeth

The rated contact stress is

$$\sigma_H = 6.7 \times 10^4 \sqrt{\frac{K_{HV} K_{H\beta} T_1}{u_r d_{e1}^3 \vartheta_H}} \le [\sigma]_H .$$

Bending Stress Analysis of Wheel Teeth

The bending stress in the teeth of the spur is

$$\sigma_{F2} = \frac{2.72 \times 10^3 K_{FV} K_{F\beta} T_1 Y_{FS2}}{b d_{e1} m_e \, (m_{te}) \, \vartheta_F} \le [\sigma]_{F2} .$$

For gearings with circular teeth the module m_e is substituted for the module m_{te} in this formula. The bending stresses in the teeth of the pinion are

$$\sigma_{F1} = \sigma_{F2} Y_{FS1} / Y_{FS2} \le [\sigma]_{F1} .$$

The values of the factors Y_{FS1} and Y_{FS2}, considering tooth form and stress concentration, are taken from Table 6.14 and depend on the coefficient of displacement and the given number of teeth

$$z_{v2} = z_2 / \left(\cos^3 \beta_n \cos \delta_2 \right) ,$$

$$z_{v1} = z_1 / \left(\cos^3 \beta_n \cos \delta_1 \right) .$$

For the checking strength analysis of teeth under the action of peak loads see Sect. 6.3.7.

6.5 Worm Gearings

6.5.1 Background

Worm gearings are used for transmission of rotational motion between shafts, the axes of which intersect in space. In most cases, the intersection angle is 90° (Fig. 6.49). The drive is worm 1, representing a gear wheel with a small number ($z_1 = 1–4$) of teeth (coils), which is similar to an acme screw or an approximate thread. To increase the contact line length in the toothing with the worm, the teeth of worm wheel 2 have the form of an arc in axial section. The worm gearing is a tooth-screw gear, the motion of which is transformed according to the principle of the screw pair with its inherent increased slip [6.38–47].

Depending on the form of the external worm surface gearings can have a cylindrical worm (Fig. 6.49a) or a globoidal worm (Fig. 6.49b). The quality of globoidal gears is higher, but they are complicated to manufacture and assemble, and are sensitive to the axial displacement caused by, e.g., wear of the bearings. In practice, gearings with cylindrical worms are most often applied.

The advantages of worm gearings are:

1. The availability of a high gear ratio u in one stage (up to 80).
2. Compactness and moderate mass of the structure.
3. Operation smoothness and silence.
4. The availability of self-stopping gearings, i. e., permitting motion only from the worm to the wheel. This self-stopping of the worm gearing allows a mechanism without a braking device, preventing reverse rotation of the wheels (e.g., under the action of a lifted load gravity).
5. The availability of exact and slight displacements.

Their disadvantages are:

1. A relatively low efficiency factor because of increased slip of the worm coils on the wheel teeth, and as a result considerable heat release in the toothing zone
2. The need for expensive antifriction materials for the ring of the worm wheels
3. Increased wear and tendency to seizing
4. The necessity for adjustment of the mesh (the mean plane of the worm wheel ring must coincide with the axis of the worm)

Worm gearings are widely used in vehicles, lifting-and-shifting machines with low and mean capacity (the lifting mechanism of elevators, winch, power hoist;

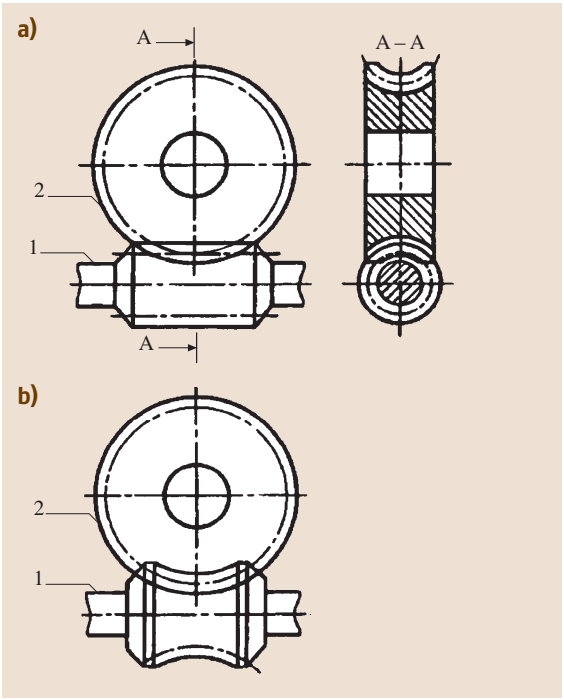


Fig. 6.49a,b Worm gear with a cylindrical **(a)** and a globoidal **(b)** worm

drives of underground escalators, transmissions of vehicles, etc.), as well as those vehicles having the purpose of slight and exact displacements (dividing devices of the machines, tuning and adjusting mechanisms, etc.).

Owing to the disadvantages mentioned above it is impractical to use worm gearings in environments of continuous action with power of more than 30 kW. By operating in intermittent cycles they can be effective also with higher power.

6.5.2 Geometry of Worm Gearings

The quality and efficiency of worm gearings depend on the form, hardness, roughness, and manufacturing accuracy of the screw surface of the worm coil. There are linear and nonlinear worms, depending on whether screw surfaces of the worm coils can or cannot be produced by a straight line. Cutting of the linear screw surfaces is carried out with universal screw-cutting lathes, in which the linear cutting edge reproduces evolvent, convolute, or Archimedean surfaces. A nonlinear screw surface is obtained by using milling cutters of bevel or toroidal form. In accordance with this fact there are involute, Archimedean, convolute, and nonlinear worms. The kind of screw surface of the worm coils obtained depends on the cutting method.

The involute worm (ZI – designation according to [6.44]) is obtained in the case of installation of the linear cutting edge in the plane tangential to the base cylinder with diameter d_b (Fig. 6.50). The left and right coil sides are cut, respectively, with cutters 1 and 2 mounted above and below the axis of the worm (see also sections C–C and B–B). In the face section (perpendicular to the axis of the worm) the coil profile of the worm is outlined with an involute surface; in axial section (A–A) it is curvilinear (convex). The involute worm represents a helical wheel of involute profile with a teeth number equal to the turn number of the worm and with a large tilt angle of the teeth. In this regard more efficient methods are involute worm cutters with worm milling and hob cutters assigned for the production of involute oblique teeth of cylindrical gearings.

To obtain high surface hardness of the turns, and thus increase the quality characteristics of the gearings, heat treatment with subsequent grinding of the work coil surfaces is used. Involute worms can be ground with high accuracy with a flat surface of the grinding wheel. Efficient cutting methods and ease of grinding determine high fabricability of the involute worms.

The Archimedean worm (ZA) is obtained by arranging that the position of the cutting edges in the plane passes through the axis of the worm. Archimedean worms have a linear profile with an angle of 2α, which equals the profile angle of the cutter, in axial section. In the face section the coil profile is outlined with an Archimedean spiral. The lateral coil surfaces of Archimedean worms can be ground only with special shapes along the compound curve grinding wheel. Strengthening heat treatment and subsequent grinding are not carried out, and Archimedean worms with low hardness are used in slow-speed gearings with low demands for load-carrying ability and lifetime.

The convolute worm (ZN) is obtained by placement of the cutting edges in the plane tangential to the cylinder with diameter d_x ($0 < d_x < d_b$) and normal to the symmetry axis of the valley. Convolute worms have convex profile in axial section; in the face section the coil section is outlined with an elongated evolvent.

The disadvantages of gearings with convolute worms are the irregular shape of the polishing tools for the worms and the impossibility of precision milling cutters for teeth cutting of the worm wheels. Like Archimedean worms, gearings with convolute worms have limited application, in general in the short-run environment.

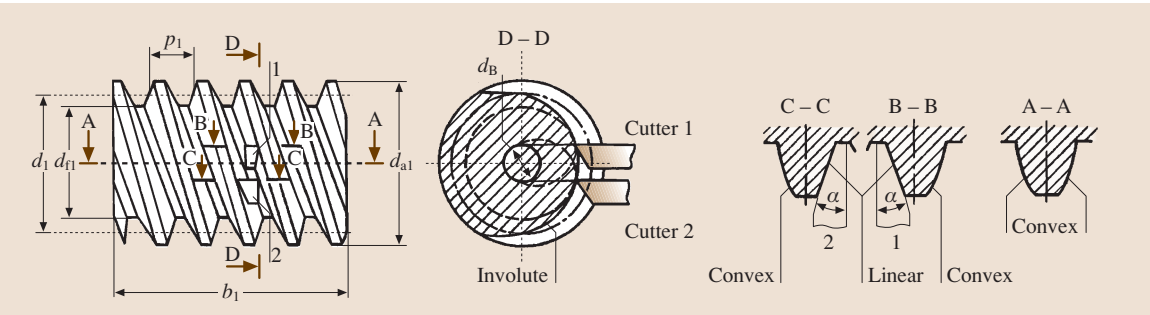


Fig. 6.50 Principal geometry of the worm

Nonlinear worms are cut with milling cutters of conical or toroidal form. The coils of such worms have a curvilinear profile in every section; in the section normal to the symmetry axis of the valley they are convex; in the axial section they are concave.

The work coil surfaces of nonlinear worms are ground with high accuracy with conical or toroidal wheels. Gearings with nonlinear worms are characterized by increased load-carrying capacity. Nonlinear worms formed by the cone (ZK) or the torus (ZT) are considered promising. Involute and nonlinear worms are used in power trains.

The geometry of the worm and the wheel is determined from formulas similar to those for gear wheels. In worm gearings the axial module m of the worm is rated and equals the face module of worm wheels. The values for m (in mm) are chosen from the series: ...4, 5, 6.3, 8, etc. The basic geometry of the worm consists of the following (Fig. 6.50):

- A pitch diameter, i. e., the diameter of the cylinder of the worm, where the coil thickness equals the width of the valley

  $$d_1 = mq \,,$$

  where q is a module number in the pitch diameter of the worm or a coefficient of the worm diameter. To reduce the number of gear cutting tools required the values of q are standardized to 8, 10, 12.5, 16, and 20.
- The rated pitch of the worm: $p_1 = \pi m$;
- Coil move: $p_{z1} = p_1 z_1$, where z_1 is the coil number of the worm, taking values of 1–4;
- The angle α of thread for involute, Archimedean, and convolute worms is $\alpha = 20°$, for worms formed by the torus it is $\alpha = 22°$;
- A pitch helix angle of the coil line (see Fig. 6.51)

  $$\tan\gamma_1 = p_{z1}/(\pi d_1) = \pi m z_1/(\pi m q) = z_1/q \,.$$

For the worm in gearings with displacement it is computed separately as follows:

- The diameter of the pitch cylinder (starting diameter)

  $$d_{w1} = m(q + 2x) \,,$$

- The helix angle of the coil line on the pitch cylinder

  $$\tan\gamma_{w1} = z_1/(q + 2x) \,,$$

where x is the coefficient of displacement.

The teeth of the worm wheel are mostly cut with a hob that represents a clone of the worm with which the worm wheel will be meshed. Only the milling cutter has cutting edges and a somewhat greater (by the double rim clearance dimension in the toothing) outer diameter. During cutting the work piece of the wheel and the milling cutter both move the same way as the worm wheel and the worm will during operation.

The basic geometry of worm-wheel rings is determined by its mean section. The pitch d_2, and hence also the starting d_{w2} diameter of the wheel, with teeth number z_2 (Fig. 6.52) is given by

$$d_2 = d_{w2} = m z_2 \,.$$

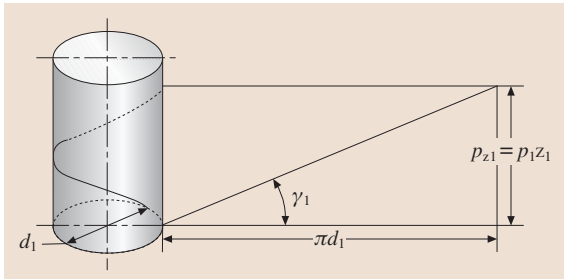


Fig. 6.51 Determination of the pitch helix angle of the coil line

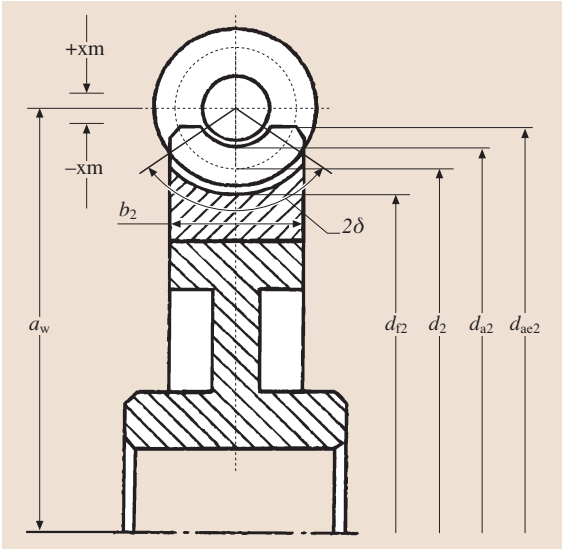


Fig. 6.52 Principal geometry of the worm wheel

The axle base of the worm gearing is

$$a = 0.5(d_1 + d_2) = 0.5(mq + mz_2) = 0.5m(q + z_2) \, .$$

Worm gearings with displacement are made in order to achieve a standard or given axle base. This is achieved in the same way as in gearings by means of a displacement by xm of the milling cutter relative to the work piece in teeth cutting of the wheel (Fig. 6.52)

$$a_w = a + xm = 0.5m(q + z_2 + 2x) \, .$$

For standard reduction gears a_w takes the values 80, 100, 125, 140, 160, etc. For teeth cutting of the wheels in gearings with and without displacement, the same tool is used. Thus, cutting with displacement is carried out only for the wheel.

With the set center distance the coefficient of tool displacement is

$$x = (a_w/m) - 0.5(q + z_2) \, .$$

The values of the tool displacement coefficient x are chosen from the data for noninterference and nonthinning of the teeth. The following is recommended for gears with worms:

- Involute: $-1 \le x \le 0$
- Formed by a torus: $0.5 \le x \le 1.5$

The worm wheel is helical, with an angle γ_w of tooth dip.

The relative angle 2δ of contact for strength analysis is determined according to the crossing points of the circle with diameter $(d_{a1} - 0.5m)$ and the end lines of the worm-wheel face.

To ensure correct toothing of worms with the wheel in a single-part environment, mean plane position control of the wheel ring relative to the worm axis is required; in a mass-manufacturing environment deliberate low surface distortion (retraction) of the wheel tooth along the depth in the direction of its point and root (profile modification), or along the length towards its ends (longitudinal modification), is resorted to.

6.5.3 The Kinematics of Worm Gearings

The gear ratio u of worm gearings is determined by the observation that, for each worm revolution, the wheel turns by an angle corresponding to the teeth number of the wheel, which equals the coil number of the worm.

The wheel turns completely during z_2/z_1 worm revolutions, so

$$u = n_1/n_2 = d_2 \operatorname{cotan} \gamma_1/d_1 = z_2/z_1 \, ,$$

where n_1 and n_2 are the rotational frequencies of the worm and the wheel, d_1 and d_2 are the pitch diameters of the worm and the wheel, γ_1 is the pitch helix angle of the coil line, and z_1 and z_2 are the coil number of the worm and the teeth number of the wheel, respectively.

To avoid interference of the tooth root $z_2 \ge 26$ is allowed in teeth cutting. $z_2 = 32$–63 is optimal. For worm gearings of standard reduction gears the gear ratios are chosen from the series: ... 31.5, 40, 50, 63, and 80.

6.5.4 Slip in Worm Gearings

In operation, the worm coils in worm gearings slip on the teeth of the worm wheel. The slip velocity v_{sl} (Fig. 6.53) is directed at a tangent to the coil line of the worm and can be determined from the parallelogram of velocities (v_1 and v_2 are circumferential velocities of

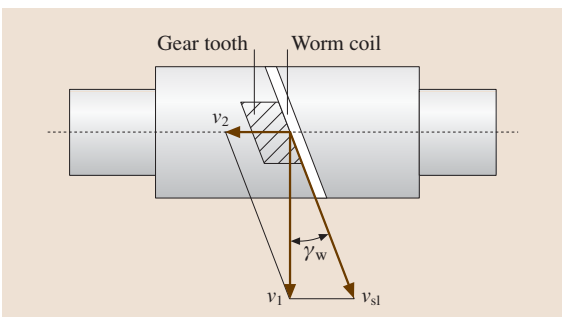


Fig. 6.53 Determination of the slip velocity in worm gears

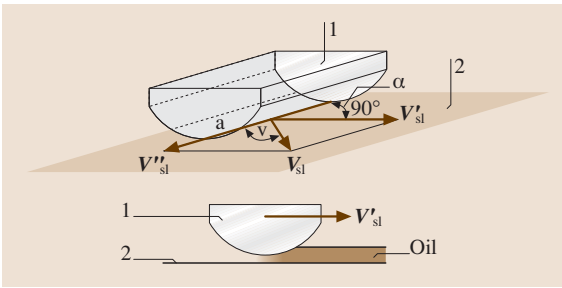


Fig. 6.54 Formation of the base oil layer in contact

the worm and the wheel, m/s) as

$$v_{sl} = v_1 / \cos\gamma_w = \pi d_{w1} n_1 / (60\,000 \cos\gamma_w) .$$

Obviously $v_{sl} > v_1$. Slip in worm gearings causes considerable losses in toothing, gearing heat, and wear of the worm-wheel teeth, and increases the seizing inclination.

The load-carrying capacity of movement on lubricated surfaces is considerably higher if wedge clearance is ensured in the direction of velocity. For surfaces with line contact, e.g., bodies 1 and planes 2 (Fig. 6.54) this corresponds to the condition that the velocity vector v_{sl} is perpendicular to the line contact a–a, or that it has the constituent v'_{sl}, which is perpendicular to that line. At the same time, oil sucked into the wedge clearance separates the mated surfaces and transfers the active load. If slip increases along the contact line a–a, the oil layer in the contact zone cannot be formed and conditions for seizing occur.

In worm gearings the contact line configuration depends on the form of the worm surface. In Fig. 6.55 the layout of the worm-wheel tooth has the contact line a–a embossed on it; the series of subsequent positions is

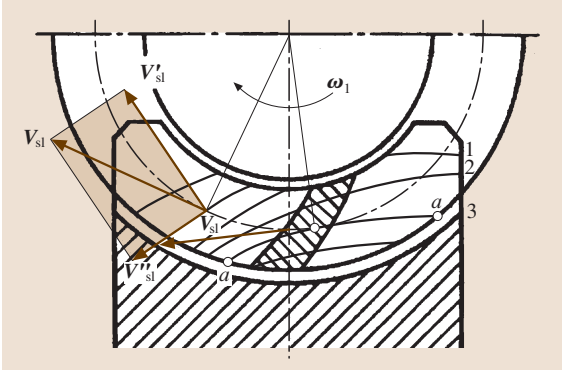


Fig. 6.55 Position of the contact lines on the tooth of the worm wheel

marked with the numbers 1, 2, and 3 in the toothing process of wheels with Archimedean worms. Obviously, in position 3 the slip velocity vector v_{sl} is directed at a tangent to the contact line. The area where the direction v_{sl} almost coincides with the direction of the contact lines is shaded. An unfavorable direction of the slip velocity is the cause of increased losses, and wear and seizing starting exactly in this area and are then distributed throughout the work surface of the wheel teeth.

The most favorable contact zone is the part of the wheel tooth from the side of the worm lead-out of the toothing. Here the vector v_{sl} has a considerable constituent v'_{sl}, which is perpendicular to the contact line, and therefore there are favorable conditions for the formation of the base oil layer.

In worm gearings with nonlinear worms the contact lines are located so that in all positions a considerable constituent v'_{sl} of the slip velocity vector takes place in the toothing process. This provides increased load-carrying capacity for the gearing.

Twelve accuracy degrees are determined for worm gearings. Kinematic accuracy standards, smoothness standards, and contact standards of the teeth and the coils are provided for all of these. In power trains the seventh ($v_{sl} \le 10$ m/s), eighth ($v_{sl} \le 5$ m/s), and ninth ($v_{sl} \le 2$ m/s) accuracy degrees are mostly used.

6.5.5 The Efficiency Factor of Worm Gearings

Lubrication in worm gearing is more important than in toothed gears because coil slipping of the worm occurs along the contact lines of the worm-wheel teeth in toothing. The worm gearing is a tooth-screw and has losses that are inherent to the toothed gear as well as to the screw-nut gear. Generally the efficiency factor of the worm gearing is

$$\eta = \eta_b \eta_t \eta_{st} ,$$

where η_b, η_t and η_{st} are efficiency factors that take into account losses in bearings, toothing, and in oil stir and splatter, respectively.

The efficiency factor of the worm meshing is determined from the formula

$$\eta_t = \tan\gamma_w / \tan(\gamma_w + \rho) ,$$

where γ_w is the helix angle of the screw line and ρ is the modified angle of friction.

The values of the friction angle ρ, depending on the slip velocity, are obtained experimentally for worm

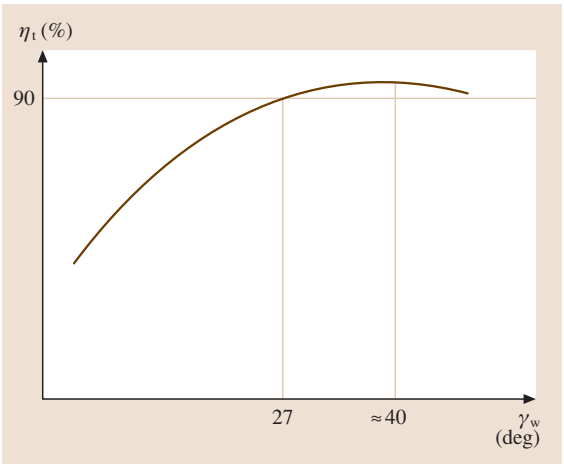


Fig. 6.56 Dependence of the efficiency factor η_t on the helix angle γ_w of the coil for the pitch cylinder of the worm

gearings on supports with frictionless bearings; i. e., capacity losses in frictionless bearings, in toothing, and in oil stir and splatter are considered in these values. The value r reduces if v_{sl} increases, because with high slip velocities in the contact area favorable conditions for oil layer formation occur, which separates the worm coils and wheel teeth and reduces losses in the toothing.

The numerical value of η_t increases with the helix angle γ_w increasing the screw line on the pitch cylinder up to $\gamma_w \approx 40°$ (Fig. 6.56). Usually, in worm gearings one has $\gamma_w \leq 27°$. Large helix angles are applied in gearings with four-thread worms and small gear ratios. Worm gearings have a relatively low efficiency factor, which limits their application field to $\eta_t = 0.75–0.92$.

6.5.6 Toothing Forces

The interaction force between the worm and the wheel is taken to be concentrated and is applied at the pitch point along the normal line to the work surface of the coil. It is described by three mutually perpendicular constituents: F_t, F_a, and F_r. For clarity of the force representation, the worm and worm wheel in Fig. 6.57a have been disengaged.

The peripheral force F_{t2} on the worm wheel is

$$F_{t2} = 2\times10^3 T_2/d_2 \,,$$

where T_2 is the torque on the worm wheel (N m) and d_2 is the pitch diameter of the wheel (mm).

The axial force F_{a1} on the worm numerically is F_{t2}, i.e., $F_{a1} = F_{t2}$.

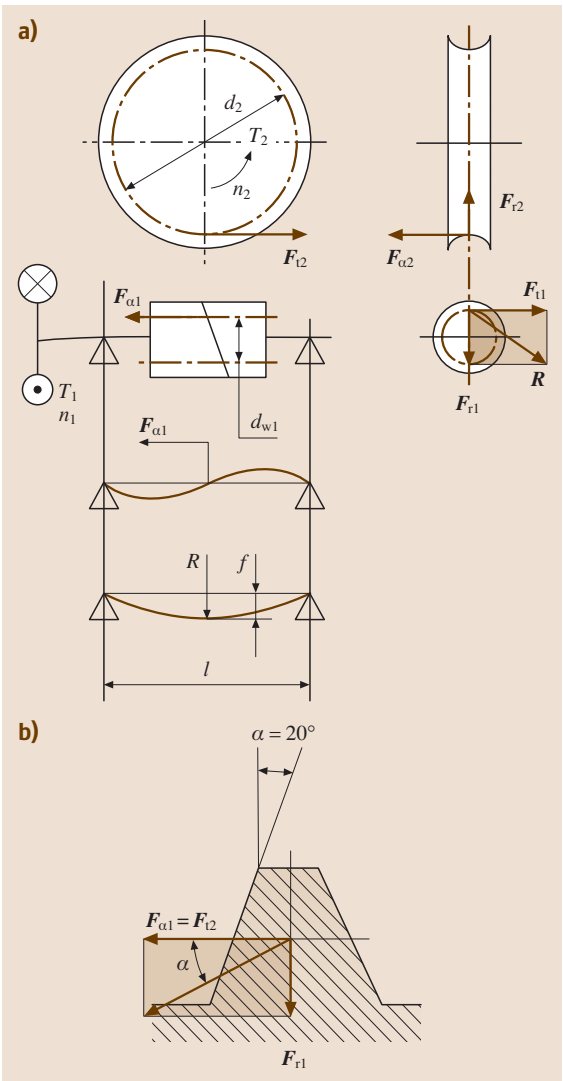


Fig. 6.57a,b Forces in the toothing of worm gears (a), layout to determine F_{r1} (b)

The peripheral force F_{t1} on the worm is

$$F_{t1} = 2\times10^3 T_1/d_{w1} = 2\times10^3 T_2/(u\eta_t d_{w1}) \,,$$

where T_1 is the torque on the worm (N m), η_t is the efficiency factor of the gearing, and d_{w1} is in mm.

The axial force F_{a2} on the worm wheel numerically equals F_{t1}, i.e., $F_{a2} = F_{t1}$. The radial force F_{r1} on the worm (the radial force F_{r2} on the wheel numerically equals F_{r1}) is (Fig. 6.57b)

$$F_{r1} = F_{r2} = F_{t2} \tan\alpha / \cos\gamma_w \,.$$

The direction of the force F_{t2} always coincides with the sense of rotation of the wheel, and the force F_{t1} is directed towards the side opposite the worm rotation.

6.5.7 Stiffness Testing of Worms

Hardness analysis of the worm body under the action of the forces in the toothing is carried out in order to prevent excessive load concentration in the contact area.

Flexing due to the constituents F_{t1} and F_{r1} in the worm bearing section, where the most important toothing area is located, have maximum values. Flexing in this section due to the torque produced by the axial force F_{a1} is zero (Fig. 6.57a).

Flexing f of the worm due to the resultant radial force R leads to an increase of the axle base and an increase of the pitch cylinder radius of the worm. The angle of the coil dip on the deformed worm does not equal the angle of the teeth dip of the worm wheel; toothing precision is broken, which causes load concentration in the toothing.

Thus, worm flexing f (mm), in the mean section is limited to the allowed values $[f] = (0.005–0.008)m$ (where m is the toothing module, mm)

$$f = \frac{\sqrt{F_{t1}^2 + F_{r1}^2} \times l^3}{48 E J_e} \leq [f] \ ,$$

where l is the distance between the worm bearings (mm) (in predesigns, $l \approx 0.9 d_2$ can be used), E is the coefficient of elasticity of the worm material (N/mm^2), J_e (mm^4) is the equivalent moment of inertia (the moment of inertia of the cylindrical bar, which is distortion equivalent to the worm),

$$J_e = \frac{\pi d_{f1}^4}{64} \left(0.36 + 0.64 \frac{d_{a1}}{d_{f1}}\right) \ .$$

6.5.8 Materials for Worms and Worm-Wheel Rings

Worms and wheels must have sufficient strength and form a well-ground antifriction pair in view of the considerable slip velocities in the toothing. Worms are manufactured from medium-carbon steels of grades C 46, C 45 (EN), C 53, and C 50 E (EN) or alloy steels of grades 37 Cr 4, 41 Cr 4 (EN), and 40 NiCr 6 (DIN) with surface hardening or volume quenching up to hardnesses of 45–54 HRC and subsequent grinding of work coil surfaces. Worms from cemented steels of grades 20 MnCr 5 G (DIN) and 20 CrS 4 (DIN) with hardness after quenching of 56–63 HRC ensure good operation (Appendix 6.A Table 6.95).

Materials for worm-wheel rings can be classified into three groups according to decreasing score-resistance and antifriction behavior, as recommended for application slip velocities (Appendices 6.A Table 6.95, Table 6.97).

Group I

Tin bronze is applied for high slip velocities ($v_{sl} = 5–25$ m/s). This material has good score resistance, but low strength.

Group II

Tinless bronze and brass are used with intermediate slip velocities ($v_{sl} = 2–5$ m/s). Most often aluminum bronze is applied. This bronze has high mechanical strength, but low score resistance, so it is used together with quenched (> 45 HRC) ground and polished worms.

Group III

Grey iron of grades ISO 150 and ISO 200 are applied for low slip velocities ($v_{sl} < 2$ m/s) in hand-driven devices.

6.5.9 The Nature and Causes of Failure of Worm Gearings

In gearings with wheels made of tin bronze (a soft material) fatigue spalling of the work surfaces of the wheel teeth is the most dangerous failure mode, because of the increasing contact stress and increasing fatigue limit of metal for the given loading cycle number.

Seizing is also possible as a result of the considerable slip velocities of the contact surfaces in combination with the boundary lubrication rate (lack of a separating oil layer). Seizing of soft materials is shown as a *smearing* (diffusion transfer) of bronze on the worm; the teeth section decreases gradually, but the gearing continues to operate for some time, determined by the wear rate.

Seizing in gear rings made of tinless bronze, brass, and iron (a hard material) results in the formation and subsequent fracture of a microwelded bridge with a jump of the friction coefficient and catastrophic wear, which results in wheel teeth damage with scales after microwelding onto the worm coils. To prevent seizing it is recommended that surfaces of the coils and teeth be thoroughly treated, and that materials with high antifriction behavior and oils with load-carrying and antiscoring additives be used.

Wear of the wheel teeth of worm gearings is mostly caused by work surface seizing of the worm/worm-wheel pair. It can also occur as a result of insufficient coil smoothness of considerably harder worms or insufficient oil purity. Higher wear leads to the formation of inadmissible clearance in the toothing. Fracture of the worm-wheel teeth occurs mostly after wear. Tooth stripping occurs very rarely.

6.5.10 Contact Strength Analysis and Seizing Prevention

For worm gearings, similarly to tooth gearings, contact strength and bending strength analysis of the worm-wheel teeth are carried out. In worm gearings, besides flaking of the work tooth surfaces, there is a risk of seizing, which also depends on the values of contact stresses σ_H. Thus, for all worm gearings contact stress analysis is the most important, determining the dimensions of the gearings, while seizure-prevention calculations and bending strength analysis are used for verification.

Contact stress analysis is carried out for the toothing in the pitch point, as applied to gearings with Archimedean worms, taking into account that the toothing conditions and the load-carrying capacity of gearings with linear-base-type worms are very similar.

To obtain the rated relation, let us express the values in Hertz's formula in terms of the parameters of the worm meshing. The force F_n acting along the normal line at the contact point of the worm-wheel tooth and the worm coil is determined according to the peripheral force F_{t2} considering the dip angle γ_w of the wheel tooth and the load factor K according to

$$F_n = KF_{t2}/(\cos\alpha\cos\gamma_w) .$$

Here the load factor $K = K_{H\beta}K_{HV}$ takes into account the unevenness of the load distribution in the contact area ($K_{H\beta}$) due to distortion of the worm and wheel shaft, the bearings, and the case, as well as the internal gearing dynamics (K_{HV}) caused by manufacturing errors.

b in Hertz's formula is understood as $l_\sum$, the total length of the contact lines in the toothing of the worm gearing. The tooth width b' of the wheel along the circular arc with diameter d_{w1} can be written in terms of the starting diameter of the worm d_{w1} and the relative contact angle 2δ (Fig. 6.52) as

$$b' = \pi d_{w1} 2\delta/360^\circ .$$

To take into consideration that, with an increase of the helix coil angle γ_w, the length of the contact line rises inversely with $\cos\gamma_w$, with the front contact ratio ε_α in the mean plane of the worm wheel and the coefficient ξ of the full-length mode of the contact line we have

$$l_\sum = b'\xi\varepsilon_\alpha/\cos\gamma_w = \frac{\pi d_{w1}}{\cos\gamma_w}\frac{2\delta}{360^\circ}\xi\varepsilon_\alpha .$$

where the mean values of the coefficient is $\xi = 0.75$, the angle of contact is $2\delta \approx 100^\circ$, and the coefficient $\varepsilon_\alpha = 2$, so the total length of the contact lines equals $l_\sum \approx 1.3 d_{w1}/\cos\gamma_w$.

The coils of Archimedean worms have a straight-sided profile ($\rho_1 = \infty$) in axial section, and the teeth of the worm wheel have an involute profile (Fig. 6.58). The radius of profile curvature of the worm-wheel tooth in the pitch point as a function of the dip angle γ_w of the tooth is

$$\rho_2 = 0.5 d_2 \sin\alpha/\cos^2\gamma_w .$$

Then

$$\sum(1/\rho_i) = 1/\rho_1 + 1/\rho_2 = \frac{1}{\rho_2} = \cos^2\gamma_w/(0.5 d_2 \sin\alpha) .$$

For the combination steel–bronze or steel–iron we have: $E_1 = 2.1\times10^5$ N/mm^2, $E_2 = 0.9–1.0\times10^5$ N/mm^2, $\nu_1 = 0.3$, and $\nu_2 = 0.25–0.33$. We assume $\alpha = 20^\circ$ and $\gamma_w = 10^\circ$.

Substituting the above parameters into the basic relation for σ_H and at the same time replacing the

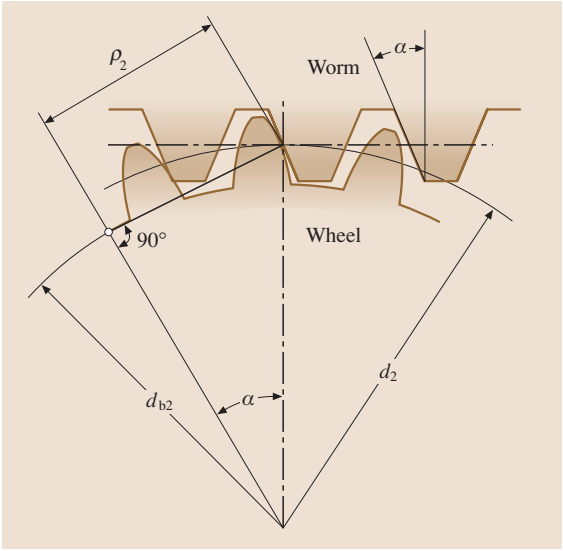


Fig. 6.58 Diagram for the contact strength analysis of the worm gear

values $F_{t2} = 2\times10^3 T_2/d_2$; $d_{w1} = m(q+2x)$; $d_2 = mz_2$; $m = 2a_w/(z_2+q+2x)$, and also using a strength condition $\sigma_H \le [\sigma]_H$, we obtain

$$\sigma_H = \frac{5350\,(q+2x)}{z_2} \times\sqrt{\left[\frac{z_2+q+2x}{a_w\,(q+2x)}\right]^3 KT_2} \le [\sigma]_H\ , \qquad (6.10)$$

where σ_H is the rated contact stress in the toothing area (N/mm^2), a_w is the axle base (mm), T_2 is the torque on the wheel (N m), and $[\sigma]_H$ is the allowable contact stress (N/mm^2).

Worm gearings with nonlinear worms are characterized by a more favorable ratio of curvature radii of the worm and the wheel, as well as a greater total length of the contact lines, which leads to increased load-carrying capacity. Contact stresses in gearings with nonlinear worms can be determined approximately from (6.10) with substitution of the numerical coefficient of 5350 for the value 4340.

Assuming the worm to be rigid, one obtains $q = 0.25z_2$ and $x = 0$. Solving (6.10) for a_w we obtain the verification analysis formula for worm gearings

$$a_w \ge K_a\sqrt[3]{KT_2/[\sigma]_H^2}\ ,$$

where $K_a = 610$ for linear worms and $K_a = 530$ for nonlinear ones, a_w is in mm, $[\sigma]_H$ is in N/mm^2, and T_2 is in N m. Substituting the parameters of worm gearings into the initial dependence for σ_H, we obtain a formula for the verification analysis

$$\sigma_H = \frac{98Z_E\cos\gamma_w}{d_2}\sqrt{\frac{KT_2}{d_{w1}\xi}} \le [\sigma]_H\ , \qquad (6.11)$$

where σ_H is a rated contact stress (N/mm^2), $Z_E = \sqrt{1/\left\{\pi[(1-\nu_1^2)/E_1 + (1-\nu_2^2)/E_2]\right\}}$ is a coefficient taking into account the stress–strain properties of the worm and worm wheel, (N/mm^2)$^{0.5}$; T_2 is the torque on the wheel (N m), d_2 and d_{w1} are the pitch diameter and the initial diameter, respectively, of the wheel and the worm (mm). ξ is a coefficient that takes into account the influence of the worm gear class on the load-carrying capacity; for linear worms one uses $\xi = 1$, whereas for nonlinear worms $\xi = 1.06 + 0.057\nu_{sl}$ on the condition that $\xi \le 1.65$, and $[\sigma]_H$ is the allowable contact stress (N/mm^2).

In short-cut calculations seizing prevention is provided in contact stress analysis by the choice of the allowable stress. In more precise calculations the worst case is supposed, i. e., when the load-carrying capacity reduction of the oil film results in immediate surface seizing. The critical temperature ϑ_{cr} of oil film breakdown has been determined experimentally for the main oil grades ($\vartheta_{cr} = 100$–$350\,^\circ$C).

The criterion for lack of seizing is represented in the form

$$\vartheta_\sum = (\vartheta + \vartheta_{mom}) < \vartheta_{cr}\ ,$$

where ϑ is the temperature of the friction surface before the contact (the oil temperature in the reduction gear), ϑ_{mom} is the instantaneous temperature on contact (*temperature flash*), which can be determined from a special calculation by solving the differential thermal conductivity equation while taking into account the specific characteristics of the behavior of the thermal process during the contact. To achieve the total temperature $\vartheta_\sum$ the critical value ϑ_{cr} can be determined experimentally.

6.5.11 Bending Strength Calculation for Wheel Teeth

This calculation is carried out for the teeth of the worm wheel, because the coils of the worm are considerably tougher. A bending calculation is performed according to the formulas for helical wheels, writing included values in terms of the parameters of the worm gearing and taking into consideration the greater teeth bending strength of worm wheels due to their arched form (Fig. 6.52).

Taking into account these features, we obtain the formula for checking the bending stress analysis of worm-wheel teeth

$$\sigma_F = \frac{KF_{t2}Y_{F2}\cos\gamma_w}{1.3\text{m}^2\,(q+2x)} \le [\sigma]_F\ , \qquad (6.12)$$

where σ_F is the design bending stress in the weakest section of the tooth (N/mm^2), Y_{F2} is the form factor of the wheel tooth, chosen depending on the equivalent tooth number z_{v2} (where greater values correspond to lower values of the tooth number), and $[\sigma]_F$ is the allowable bending stress for the wheel teeth (N/mm^2).

The *equivalent tooth number* z_{v2}, similarly to a helical wheel with dip angle γ_w, of the tooth becomes

$$z_{v2} = z_2/\cos^3\gamma_w\ .$$

6.5.12 Choice of Permissible Stresses

Permissible stresses are determined from empirical formulas depending on the material of the wheel teeth, the

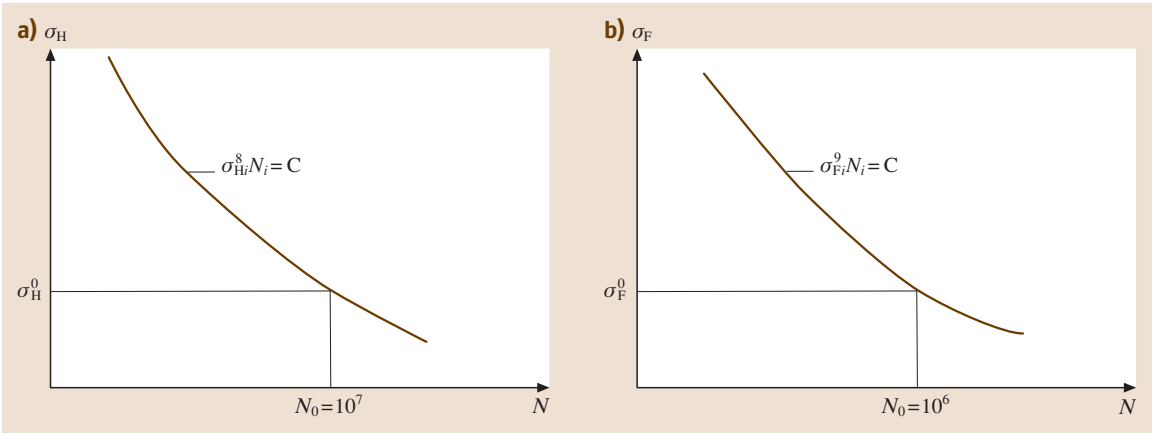


Fig. 6.59a,b Stress-cycle diagrams for (a) contact stresses and and (b) bending stresses

hardness of the worm coils, the slip velocity, and the required lifetime.

Stress-cycle diagrams for contact stresses (Fig. 6.59a) and bending stresses (Fig. 6.59b) of bronze specimens have very long inclined areas: up to 25×10^7 loading cycles. In connection with low rotational frequencies the equivalent loading cycle number of the worm-wheel teeth is small. Thus, the initial stresses are σ^0_H and σ^0_F, which equal the fatigue strength at for 10^7 and 10^6 loading cycles, respectively. For tinless bronze and iron the inclined areas of the stress-cycle diagrams are short, which allows stresses to be chosen irrespective of the cycle number. More details on the choice of permissible stresses are given in Sect. 6.5.14.

6.5.13 Thermal Design

Worm gearings operate with high levels of heat generation as a consequence of their low efficiency factor. Oil heats up to a temperature exceeding the permissible value $[t]_{oi}$, which results in a reduction of its protective ability, a breakdown of the oil film, and the possibility of seizing in the gearing.

The power $(1-\eta)P_1$ that is wasted in the toothing and due to bearing friction, as well as due to oil stir and splatter, is converted into heat, which warms the oil, the gearing units, and the walls of the case, through which it is exported to the environment.

The thermal design of worm gearings under a set operating mode (Fig. 6.60) is performed on the basis of heat balance, i. e., the equality of heat generation Q_{fr} and heat exchange Q_{sn}.

The heat flow (heat power) W of the gearing per second is

$$Q_{fr} = 10^3(1-\eta)P_1 \,,$$

where η is the efficiency factor of the worm gearing and P_1 is the power passing through the worm (in kW)

$$P_1 = T_2 n_2/(9550\eta) \,,$$

where T_2 is in (N m) and n_2 is in min^{-1}.

The heat flow (heat exchange power) W of the outer surface of the reduction gear case per second is

$$Q_{sn} = K_t(t_{oi} - t_0)A(1+\psi) \,,$$

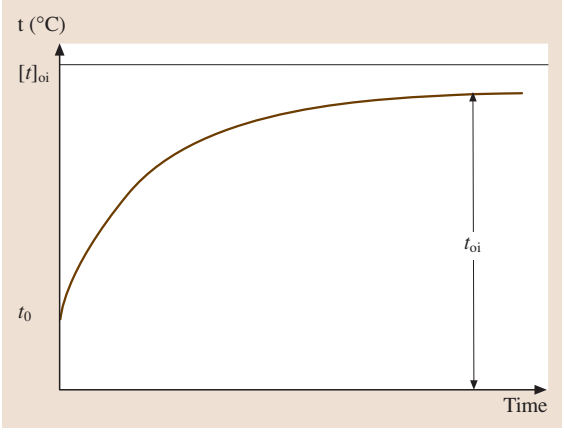


Fig. 6.60 Change in the oil temperature t_{oi} in the steady operating mode

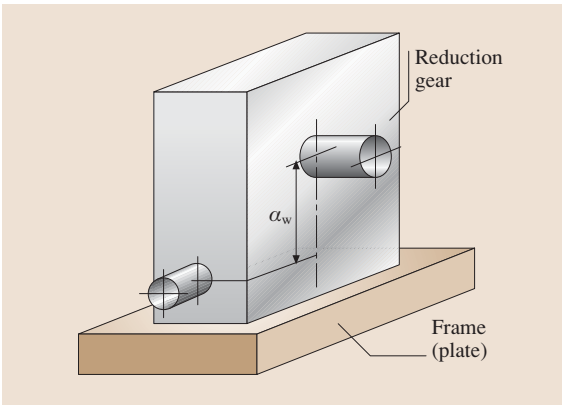


Fig. 6.61 Installation of the reduction gear on the frame (plate)

where A is the case surface area covered inside by oil or its splatter and outside by air (m^2). The bottom surface of the case is not taken into account, because air cannot freely circulate outside this area. The surface area of the cooling case A can be approximately assumed as depending on the axle base a_w (m) according to

$$A = 12a_w^{1.71} .$$

ψ is the coefficient of heat rejection from the bottom of the reduction gear to the base of the case. When mounting the reduction gear onto a metal plate or frame (Fig. 6.61), $\psi = 0–0.3$, depending on the fit of the case to the plate or frame, t_0 is the air temperature out of the case (in workshop conditions usually $t_0 = 20\,°C$), t_{oi} is the oil temperature in the case of the gearing, (°C), and K_t is the heat transfer factor defining heat flow, i.e., the heat that is transferred per second through a surface area of $1\,m^2$ for a temperature difference of 1 °C, which depends on the material of the reduction

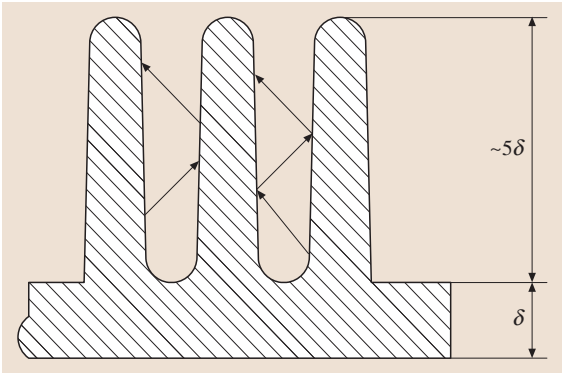


Fig. 6.62 Heat exchange between adjacent fins

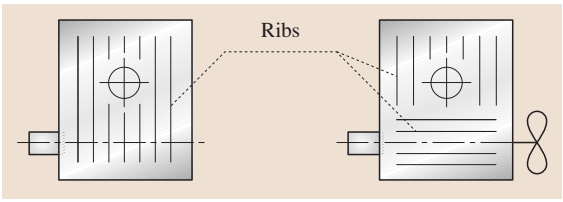


Fig. 6.63 Position of the cooling fins

gear case and the air circulation rate, i. e., the ventilation rate of the room. In the case of self-cooled cast iron one has $K_t = 12–18\,W/(m^2\,°C)$. Higher values are assumed in the case of slight roughness and clean surfaces of the exterior walls, good air circulation around the case, and intensive oil stir (due to a lower worm position).

According to heat balance condition: $Q_{fr} = Q_{sn}$, i. e.,

$$10^3(1-\eta)P_1 = K_t(t_{oi} - t_0)A(1+\psi) ,$$

hence the oil temperature t_{oi} in the case of worm gearings operated continuously without artificial cooling is

$$t_{oi} = t_0 + 10^3(1-\eta)P_1/[K_t A(1+\psi)] \le [t]_{oi} . \quad (6.13)$$

The value $[t]_{oi}$ depends on the oil grade $[t]_{oi} = 95–110\,°C$. If calculation results in $t_{oi} > [t]_{oi}$, it is necessary to increase the cooling surface A by providing cooling fins (Fig. 6.62). The fins are located upright (Fig. 6.63a), in the line of free circulated air. In calculations only 50% of the fin surface is taken into account because of the heat exchange between adjacent fins (Fig. 6.62). Artificial cooling can also be applied, e.g., using air blown onto the case with the help of a fan mounted on the worm shaft (Fig. 6.63b). The fins are positioned horizontally along the air flow direction from the fan. In this case

$$t_{oi} = t_0 + \frac{10^3(1-\eta)P_1}{[(0.65(1+\psi)K_t + 0.35K_{tv})\,A]} \le [t]_{oi} , \quad (6.14)$$

where K_{tv} is the heat transfer factor due to the air from the fan; with rotational frequencies of the worm shaft of $1000–3000\,min^{-1}$ its value falls in the range $K_{tv} = 21–40\,W/(m^2\,°C)$. In worm gearings with strong heat generation, cooling of oil with water going through the worm pipe is applied or a circulation lubrication system with a special refrigerator is used.

6.5.14 Projection Calculation for Worm Gearings

The basic data for this calculation is as follows: T_2 is the torque on the wheel (N m), n_2 is the rotational frequency of the wheel (min^{-1}), u is the gear ratio, and L_h is the service lifetime of the gearing (h).

Worm and Wheel Materials

For worms, the same steel grades are applied as for gear wheels (Table 6.1). Quenching up to hardness ≥ 45 HRC, and grinding and polishing of the worm coils are used to obtain high-quality gearings. The most practically feasible are involute worms (ZI), and promising options are nonlinear worms formed with a cone (ZK) or a torus (ZT).

Heat refining treatment with hardness ≤ 350 HB is applied for low-power (up to 1 kW) short-term gearings. The field of application of these gearings with Archimedean worms (ZA) is reduced. For power trains, involute and nonlinear worms should be used.

Materials of worm-wheel gear rings can be classified into three groups, depending on their score-resistance and antifriction behavior and whether they can be recommended for application slip velocities (Table 6.21) (Appendices 6.A Table 6.95, Table 6.97):

Group I: Tin bronze is applied for slip velocities $v_{sl} > 5$ m/s .

Group II: Tinless bronze and brass are used for slip velocity $v_{sl} = 2–5m/s$.

Group III: Soft grey iron is applied with slip velocity $v_{sl} < 2$ m/s in hand-driven devices .

As the choice of material is concerned with slip velocity, its expected value is determined previously (in m/s) as

$$v_{sl} = 0.45 \times 10^{-3} n_2 u \sqrt[3]{T_2} .$$

Permissible Contact Stresses for Material Groups

Group I – Permissible Contact Stresses.

$$[\sigma]_H = K_{HL} C_v \sigma_H^0 .$$

Here σ_H^0 is the fatigue contact point for 10^7 stress change cycles

$$\sigma_H^0 = (0.75 \text{ to } 0.9)\sigma_t .$$

Table 6.21 Mechanical characteristics of materials used for manufacture of rings of worm wheels. Casting methods are as follows: sc – spin casting, cs – chill casting, s – sand casting (in single-part production). σ_t and σ_y are, respectively, the tensile stress and the tensile yield strength for bronze (N/mm^2), and σ_{bs} is the bending strength of iron (N/mm^2)

Group	Material		Casting technology	σ_t (N/mm^2)	σ_y (N/mm^2)
I	БрО10Н1Ф1 $v_{sl} \leq 25$ m/s	CuSn8(2.1030) (DIN)	sc	285	165
	<БрО10Ф1 $v_{sl} \leq 12$ m/s	C90700 (ASTM)	cs s	245 215	195 135
	БрО5Ц5С5 $v_{sl} \leq 8$ m/s	Rg5(2.1097) (DIN)	cs s	176 147	90 80
II	БрА10Ж4Н4 $v_{sl} \leq 5$ m/s	NiAlBz(2.0971) (DIN)	sc cs	700 587	460 430
	БрА10Ж3Мц2 $v_{sl} \leq 5$ m/s	C63200 (ASTM)	cs s	550 450	360 300
	БрА9Ж3Л $v_{sl} \leq 5$ m/s	C95200 (ASTM)	sc cs s	500 490 390	200 195 195
	ЛАЖМц66-6-3-2 $v_{sl} \leq 4$ m/s	CuZn40Mn2(2.0572) (DIN)	sc cs s	500 450 400	330 295 260
III	СЧ15, СЧ20 $v_{sl} \leq 2$ m/s	GG-15 (DIN), GG-20 (DIN)	s s	$\sigma_{bs} = 320$ N/mm^2 $\sigma_{bs} = 360$ N/mm^2	

The factor 0.9 is used for worms with hard ($H \geq 45$ HRC), ground and polished coils. The factor 0.75 is used for worms with hardness ≤ 350 HB, and σ_t is taken from Table 6.21.

The service life ratio is $K_{HL} = \sqrt[8]{10^7/N_{HE}}$, under the condition that $K_{HL} \leq 1.15$. Here $N_{HE} = K_{HE}N_k$ is an equivalent loading cycle number for the worm-wheel teeth for the whole lifetime of the gearing. If $N_{HE} > 25\times10^7$, it is assumed that $N_{HE} = 25\times10^7$.

The total cycle number of stress change is

$$N_k = 60n_2L_h\ ,$$

where L_h is the service lifetime of the gearing (h).

The values of the equivalence factors K_{HE} for typical loading conditions are given in Table 6.23.

The coefficient C_v takes the material wear rate of the wheel into account. It is assumed to depend on the slip velocity v_{sl}:

Table 6.22 The values of the coefficient C_v depending on the slip velocity v_{sl}

v_{sl} (m/s)	C_v
5	0.95
6	0.88
7	0.83
≥ 8	0.80

or in accordance with the formula $C_v = 1.66v_{sl}^{-0.352}$ on the condition that $C_v \geq 0.8$.

Group II – Permissible Contact Stresses.

$$[\sigma]_H = \sigma_H^0 - 25v_{sl}\ .$$

Here $\sigma_H^0 = 300\,N/mm^2$ for worms with hardness on the coil surface ≥ 45 HRC, whereas $\sigma_H^0 = 250\,N/mm^2$ for worms with hardness ≤ 350 HB.

Group III – Permissible Contact Stresses.

$$[\sigma]_H = \sigma_H^0 - 35v_{sl}\ .$$

Here $\sigma_H^0 = 200\,N/mm^2$ for worms with hardness on the coil surface ≥ 45 HRC, whereas $\sigma_H^0 = 175\,N/mm^2$ for worms with hardness ≤ 350 HB.

Allowable Bending Stresses

Allowable bending stresses are calculated for the teeth material of the worm wheel according to

$$[\sigma]_F = K_{FL}\sigma_F^0\ ,$$

Table 6.23 Coefficients of equivalence for the typical loading conditions of worm gearings

Typical condition	Equivalence factors	
	K_{HE}	K_{FE}
0	1.0	1.0
I	0.416	0.2
II	0.2	0.1
III	0.121	0.04
IV	0.081	0.016
V	0.034	0.004

where σ_F^0 is a fatigue bending point for 10^6 stress change cycles. For materials of groups I and II

$$\sigma_F^0 = 0.25\sigma_y + 0.08\sigma_t\ ,$$

whereas for materials of group III

$$\sigma_F^0 = 0.22\sigma_{bs}\ .$$

The service life ratio is

$$K_{FL} = \sqrt[9]{10^6/N_{FE}}\ .$$

Here $N_{FE} = K_{FE}N_k$ is an equivalent loading cycle number for the worm-wheel teeth, and N_k is the total number of stress change cycles for the whole lifetime of the gearing. If $N_{FE} < 10^6$, it is assumed that $N_{FE} = 10^6$. If $N_{FE} > 25\times10^7$, it is assumed that $N_{FE} = 25\times10^7$.

The values of the equivalence factors K_{FE} for typical loading conditions are given in Table 6.23.

Overload Stress Capacity

The overload stress capacity on the maximum static or unit peak load for materials is

Group I: $[\sigma]_{H\,max} = 4\sigma_y$; $[\sigma]_{F\,max} = 0.8\sigma_y$.
Group II: $[\sigma]_{H\,max} = 2\sigma_y$; $[\sigma]_{F\,max} = 0.8\sigma_y$.
Group III: $[\sigma]_{H\,max} = 1.65\sigma_{bs}$; $[\sigma]_{F\,max} = 0.75\sigma_{bs}$.

The Axle Base

The axle base (mm) is

$$a_w \geq K_a\sqrt[3]{K_{HV}K_{H\beta}T_2/[\sigma]_H^2}\ ,$$

where $K_a = 610$ for involute, Archimedean, and convolute worms; $K_a = 530$ for nonlinear worms. K_{HV} is a coefficient of internal dynamics, taking the value $K_{HV} = 1$ for $v_2 \leq 3$ m/s, and $K_{HV} = 1–1.3$ for $v_2 > 3$ m/s, where v_2 is the circumferential velocity of

the worm-wheel. $K_{H\beta}$ is a load concentration coefficient: under constant loading conditions $K_{H\beta} = 1$, wile under varying conditions

$$K_{H\beta} = 0.5\left(K^0_{H\beta} + 1\right) .$$

The initial load concentration coefficient $K_{H\beta}$ is found from the diagram shown in Fig. 6.64, for which the number z_1 of worm coils is determined depending on the gear-ratio u:

Table 6.24 Number of worm threads z_1 depending on gear-ratio u

u	z_1
More than 8 up to 14	4
More than 14 up to 30	2
More than 30	1

For the standard worm-and-worm pair, the calculated axle base a_w is rounded to the given value (series 1 should be preferred to series 2):

Table 6.25 Standard values of a_w

Series 1 (mm)	Series 2 (mm)
63	71
80	90
100	112
125	140
160	180
200	224
250	280
315	355

Main Parameters of Worm Gearings

The number of wheel teeth is $z_2 = z_1 u$, rounded to the nearest whole number. The tentative values are

For the module of the gearing

$m = (1.4–1.7)a_w/z_2$,

For the coefficient of the worm diameter

$q = 2a_w/m - z_2$.

The nearest standard value m to the calculated one is substituted into the formula for q (Table 6.26).

Series 1 should be preferred to series 2. The series 3 modules may be used for standardized reduction gears in machine-building applications.

Table 6.26 Standard values of module m

Series 1 (mm)	Series 2 (mm)	Series 3 (mm)
2.0	3.0	2.25
2.5	3.5	2.75
3.15	6.0	4.5
4.0	7.0	9.0
5.0	12.0	11.0
6.3		14.0
8.0		
10.0		
12.5		
16.0		

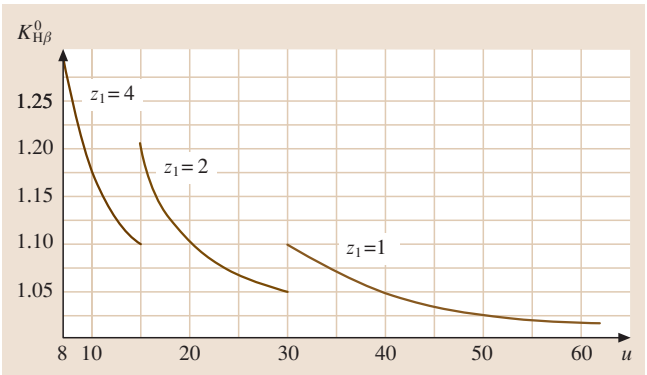


Fig. 6.64 The values of the initial concentration factor $K^0_{H\beta}$ depending on the coil number z_1 of the worm and the gear ratio u

The calculated value q is rounded to the nearest standard value (series 1 should be preferred to series 2) according to:

Table 6.27 Standard values of q

Series 1	Series 2
8.0	7.1
10.0	9.0
12.5	11.2
16.0	14.0
20.0	18.0

The minimum permissible value q from the rigidity condition of the worm is $q_{min} = 0.212 z_2$. The coefficient of displacement is

$$x = a_w/m - 0.5(z_2 + q) ,$$

which is recommended for gearings with the following worms

Involute (ZI)	$-1 \le x \le 0$,
Formed by a torus (ZT)	$0.5 \le x \le 1.5$,
Archimedean (ZA), convolute (ZN) ,	
Formed by a cone(ZK)	$0 \le x \le 1.0$.

For the helix angle of the coil worm line on the cylinder

The pitch angle	$\gamma = \arctan(z_1/q)$,
The starting angle	$\gamma_w = \arctan[z_1/(q+2x)]$,
The base angle	
(for the wormZI)	$\gamma_b = \arccos(0.940 \cos \gamma)$.

With the exception of those cases caused by drive kinematics, the worms of gearings have a coil line of the right direction. The actual gear ratio is $u_r = z_2/z_1$. The calculated value u_r must not differ from the target value by more than 4%; for standardized reduction gears for machine-building applications the tolerances are 6.3% for single gearing and 8% for double gearing .

Dimensions of the Worm and the Wheel

The dimensions of the worm and the wheel are as follows (Figs. 6.51 and 6.52):
The pitch diameter of the worm

$$d_1 = qm \ .$$

The diameter of the coil crests

$$d_{a1} = d_1 + 2m \ .$$

The diameter of the roots

$$d_{f1} = d_1 - 2.4m \ .$$

The pitch diameter of the wheel

$$d_2 = z_2 m \ .$$

The diameter of the tooth tops

$$d_{a2} = d_2 + 2m(1+x) \ .$$

The socket diameter for gearings with ZI worms

$$d_{f2} = d_2 - 2m(1 + 0.2 \cos \gamma - x) \ .$$

The socket diameter for gearings with ZN, ZA, ZK, and ZT worms:

$$d_{f2} = d_2 - 2m(1.2 - x) \ .$$

The largest diameter of the wheel

$$d_{ae2} \le d_{a2} + 6m/(z_1 + k) \ ,$$

where $k = 2$ for gearings with ZI, ZA, ZN, and ZK worms; and $k = 4$ for gearings with ZT worms.

The length b_1 of the cut worm part is

$$b_1 = 2\sqrt{(0.5d_{ae2})^2 - (a_w - 0.5d_{a1})^2} + 0.5\pi m \ .$$

The face width b_2 of the worm wheel for the gearings is:

- for ZI, ZA, ZN, and ZK worms

$$b_2 = 0.75 d_{a1} \quad \text{for } z_1 \le 3 \ ,$$
$$b_2 = 0.67 d_{a1} \quad \text{for } z_1 = 4$$

- and for ZT worms

$$b_2 = (0.7 - 0.1x) d_{a1}$$

Checking Strength Analysis of Gearings

The slip velocity in the toothing is

$$v_{sl} = v_{w1} / \cos \gamma_w \ ,$$

where $v_{w1} = \pi n_1 m(q + 2x)/60\,000$.

Here v_{w1} is the circumferential velocity on the starting diameter of the worm (m/s), $n_1 = n_2 u_r$, (min^{-1}), m is in mm, and γ_w is the initial helix coil angle. The allowable stress $[\sigma]_H$ is specified according to the rated value v_{sl}.

The design stress is determined from (6.11) by specifying the load factor as the value of $K = K_{HV} K_{H\beta}$. The circumferential velocity of the worm wheel (m/s) is $v_2 = \pi n_2 d_2/60\,000$. In the case of common manufacturing accuracy and under the condition of worm rigidity it is assumed that $K_{HV} = 1$ for $v_2 \le 3$ m/s. For $v_2 > 3$ m/s the value of K_{HV} is assumed to be equal to the coefficient K_{HV} (Table 6.8) for helical gearings with a hardness of the working tooth area ≤ 350 HB and the same accuracy degree.

The load concentration coefficient $K_{H\beta}$ is $K_{H\beta} = 1 + (z_2/\theta)^3(1 - X)$, where θ is the coefficient of worm strain (Table 6.29); X is a coefficient that takes into account the influence of the operating gearing mode on the grind of the worm-wheel teeth and the worm coils.

The values of X for typical loading conditions and cases, when the rotational frequency of the worm-wheel shaft does not change with load modification, are detailed in in Table 6.30.

Table 6.29 Deformation coefficients θ of worms

z_1	θ for coefficient q of the worm diameter					
	8	10	12.5	14	16	20
1	72	108	154	176	225	248
2	57	86	121	140	171	197
4	47	70	98	122	137	157

Table 6.30 Influence coefficients X in operating mode on running-in of worm gearings

X	Typical conditions
0	1.0
I	0.77
II	0.5
III	0.5
IV	0.38
V	0.31

The Efficiency Factor of Gearings

The efficiency factor for worm gearings is

$$\eta = \tan\gamma_w / \tan(\gamma_w + \rho) ,$$

where γ_w is a helix angle of the coil line on the pitch cylinder, and ρ is a modified friction angle determined experimentally, taking into account the relative capacity loss in the toothing, in the bearings, and due to oil stirring. The value of the friction angle ρ between a steel worm and a bronze (brass, iron) wheel depends on the slip velocity v_{sl}:

Table 6.31 Angle friction values ρ depending on the slip velocity v_{sl}

v_{sl} (m/s)	ρ	
0.5	3°10′	3°40′
1.0	2°30′	3°10′
1.5	2°20′	2°50′
2.0	2°00′	2°30′
2.5	1°40′	2°20′
3.0	1°30′	2°00′
4.0	1°20′	1°40′
7.0	1°00′	1°30′
10	0°55′	1°20′
15	0°50′	1°10′

The lower value of ρ is for tin bronze and the higher one is for tinless bronze, brass, and iron.

Toothing Forces (Fig. 6.57)

The peripheral force on the wheel, which is equal to the axial force on the worm, is

$$F_{t2} = F_{a1} = 2\times10^3 T_2/d_2 .$$

The peripheral force on the worm, which is equals to the axial force on the wheel, is

$$F_{t1} = F_{a2} = 2\times10^3 T_2/(d_{w1}u_2\eta) .$$

The radial force is

$$F_r = F_{t2}\tan\alpha/\cos\gamma_w .$$

For the standard angle $\alpha = 20°$, $F_r = 0.364F_{t2}/\cos\gamma_w$.

Bending Stress Analysis of Wheel Teeth

The calculated bending stress is determined from (6.12), where K is a load factor, the values of which are computed in the paragraph *Checking Strength Analysis of Gearing*, and Y_{F2} is the form factor of the wheel tooth, which is chosen depending on the equivalent tooth number $z_{v2} = z_2/\cos^3\gamma_w$:

Table 6.28 Tooth form coefficient Y_{F2}

z_{v2}	Y_{F2}
20	1.98
24	1.88
26	1.85
28	1.80
30	1.76
32	1.71
35	1.64
37	1.61
40	1.55
45	1.48
50	1.45
60	1.40
80	1.34
100	1.30
150	1.27
300	1.24

Checking Strength Analysis of Worm-Wheel Teeth under the Action of Peak Load

The action of the peak load is estimated by the overload factor $K_{load} = T_{max}/T$, where T is the nominal torque. The contact strength analysis under the short-term action of the maximum torque is

$$\sigma_{H\,max} = \sigma_H\sqrt{K_{load}} \le [\sigma]_{H\,max} \ .$$

The bending stress analysis of the worm-wheel teeth under the action of the maximum torque is

$$\sigma_{F\,max} = \sigma_F K_{load} \le [\sigma]_{F\,max} \ .$$

The allowable stresses $[\sigma]_{H\,max}$ and $[\sigma]_{F\,max}$ are taken from paragraph *Overload Stress Capacity*.

Thermal Design

The worm-and-wheel gearbox is checked on heating in connection with the low efficiency factor and high heat release. The power (W) on the worm is $P_1 = 0.1T_2n_2/\eta$.

The reheating temperature of oil (in the casing) under set thermal conditions without artificial cooling is determined from (6.13); with fan cooling it is determined from (6.14).

6.6 Design of Gear Wheels, Worm Wheels, and Worms

6.6.1 Spur Gears with External Toothing

The form of a gear wheel can be flat (Fig. 6.65a,b) or with a salient hub (Fig. 6.65c). Rarely (in single-reduction units), the wheel is made with a hub, which is salient on both sides [6.12, 48–56].

In Fig. 6.65 elementary forms of the wheels that are produced in *single-part small-quantity production* are shown. Wheels with small diameters are manufactured from a bar; in the case of large diameters the workpieces are produced in open forging with a subsequent turning process. To replace the requirements for precise machining with cutting, recesses are made on the wheel disks (Fig. 6.65b,c). As a rule, for diameters $d_a < 80$ mm, these recesses are not made (Fig. 6.65a).

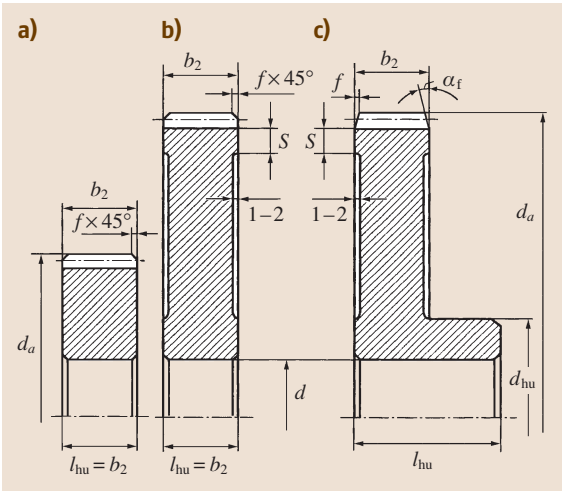


Fig. 6.65a–c Structural forms of the wheels produced in small-lot production. **(a,b)** Plane, **(a)** with a salient hub

It is advisable to assume that the length l_{hu} of the wheel slot is equal to or longer than the face width b_2 ($l_{hu} \ge b_2$). The standard hub length is adjusted by using that obtained as a result of calculation of the spline connection, with a tightness or key joint, which was chosen for torque transfer from the wheel to the shaft, and (with slot diameter d)

$$l_{hu} = (0.8–1.5)d \ , \quad \text{usually} \quad l_{hu} = (1.0–1.2)d \ ,$$

where $l_{hu} > b_2$, the projection of the hub, is positioned in the axial force direction F_a of the toothing.

The diameter d_{hu} is set depending on the material of the hub: for steel $d_{hu} = (1.5–1.55)d$, for iron $d_{hu} = (1.55–1.6)d$, and for light alloys $d_{hu} = (1.6–1.7)d$; lower values are used a spline connection of a wheel with a shaft, whereas higher values are used for a key joint and pressure coupling joint.

The cutting width S of the gear ring is assumed to be

$$S = 2.2m + 0.05b_2 \ ,$$

where m is the toothing modulus (mm).

The bevels are made on the ring ends (teeth and crown angle): $f = (0.5 - 0.6)m$, which are rounded to a standard value (see below).

On the spurs the bevel is made on-the-miter $\alpha_f = 45°$; on helical and herring-bone wheels with a working surface hardness of less than 350 HB it is made on-the-miter $\alpha_f = 45°$ (Fig. 6.65a,b), and with higher hardness $\alpha_f = 15–20°$ (Fig. 6.65c).

Pointed edges on the hub ends are also blunted with the bevels, the dimensions of which are assumed to be:

Table 6.32 Facet size f depending on d

d (mm)	f (mm)
20–30	1.0
30–40	1.2
40–50	1.6
50–80	2.0
80–120	2.5
120–150	3.0
150–250	4.0
250–500	5.0

In the case of *lot production* the workpieces of the wheels are obtained from a bar through open forging as well as through forging in dies. Yearly output of wheels of more than 50 pieces forged in elementary one-sided pad dies is economically feasible. In this case, the form of the gear wheels is designed according to the type shown in Fig. 6.66a,b.

With a yearly output of more than 100 items, double-sided dies are used. In this case, the form of the wheel is designed according to Fig. 6.67a,b. The wheel billet after die forming is shown by a fine line. For easy removal of the billet from the die, the forming gradients are assumed to obey $\gamma \geq 7°$ and the rounded radii to obey $R \geq 6$ mm. The disk thickness is

$$C \approx 0.5(S + S_{hu}) \geq 0.25 b_2 \,, \quad \text{where}$$

$$S_{hu} = 0.5(d_{hu} - d) \,.$$

For reduction of the influence of heat treatment on geometric accuracy the gear wheels are made to be massive: $C = (0.35–0.4)b_2$.

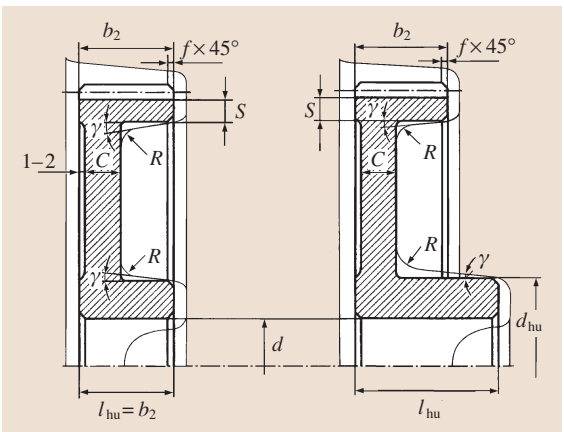


Fig. 6.66a,b Structural forms of wheels where workpieces are made through open forging or in one-sided dies. **(a)** With a nonsalient hub, **(b)** with a salient hub

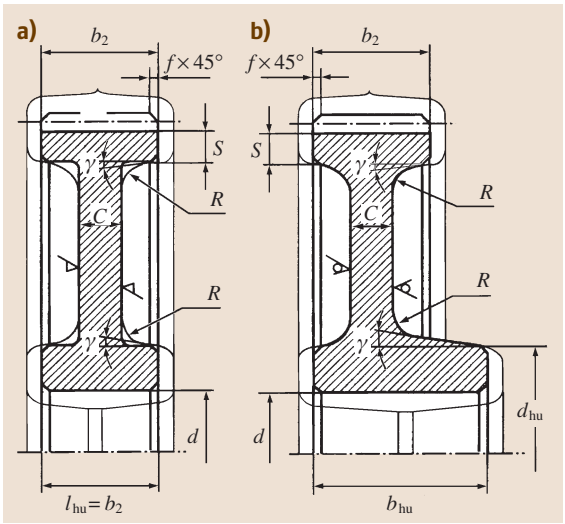


Fig. 6.67a,b Structural forms of wheels where workpieces are made in double-sided dies. **(a)** With a nonsalient hub, **(b)** with a salient hub

Plastic forming conditions of metal are improved if the grooves in the wheel disks are made as illustrated in Fig. 6.68. The rounded radii are assumed to be $R \geq 20$ mm, and the forming gradients are $\gamma \geq 12°$.

Depending on the ratio of the wheel dimensions, the grooves in the disks are shaped with one radius arc R (Fig. 6.68a) or with two arcs and a straight line segment (Fig. 6.68b). In this case, the disk thickness is $C \approx 0.5 b_2$.

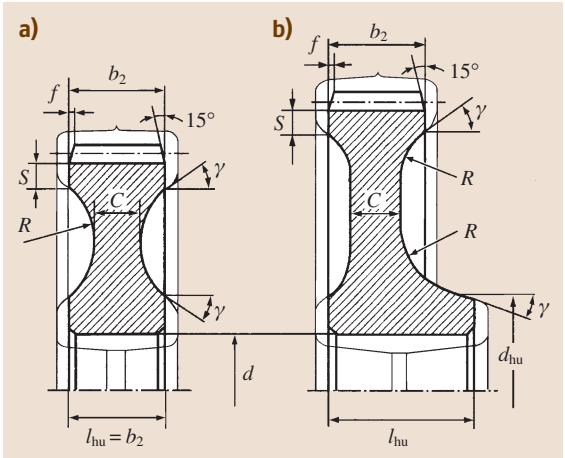


Fig. 6.68a,b Structural forms of wheels with improved straining conditions of metal due to die forming. **(a)** With a nonsalient hub, **(b)** with a salient hub

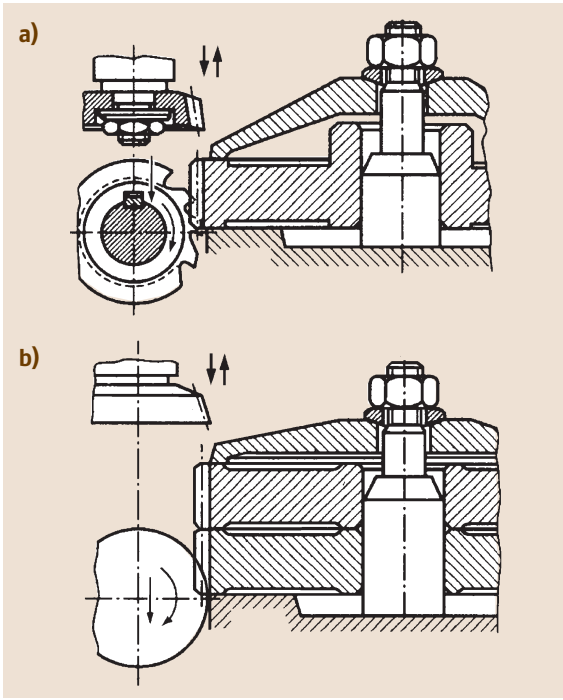


Fig. 6.69a,b Position of the workpiece for teeth cutting of one wheel (a) of the *package* of two and more wheels (b)

In automobile and aircraft construction, wheels are manufactured with a thinner disk ($C = 0.25b_2$); four to six holes of major diameter are made in the disk,

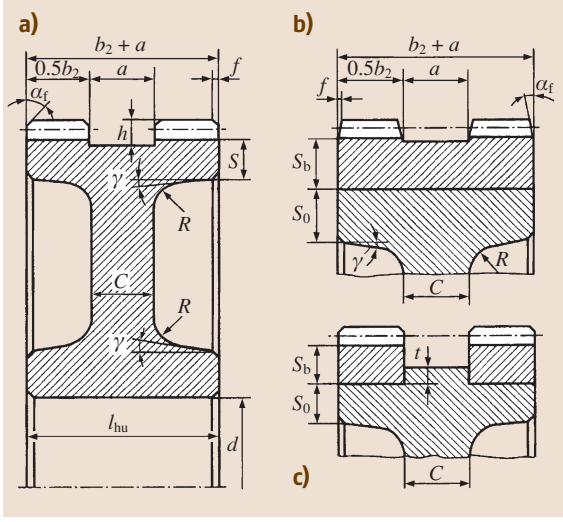


Fig. 6.70a–c Structural forms of the chevron gear. (a) One-piece, (b) and (c) composite

and the rounded radii are assumed to be minimum. Gear wheels rotating with a relatively high frequency ($n \geq 1000\,\mathrm{min}^{-1}$) are treated over their entire surface (Fig. 6.67a) and balanced.

The base surfaces in tooth *cutting* are the surface of the central bore and of the gear-ring ends. A tooth cutting layout chart for a wheel is shown in Fig. 6.69a. Output capacity increases during tooth cutting of a set of two or more wheels (Fig. 6.69b). To provide a juxtaposition of the gear-ring ends it should be mentioned in the specification that the hub end must not project beyond the gear-ring end.

Herring-bone gears (Fig. 6.70a–c) are distinguished from other cylindrical gears by their increased width. Most often herring-bone gears are manufactured with a flute in the middle, which is used for the runout of the worm hob that cuts the teeth. The width a of the flute is determined according to the diameter of the cutter depending on the module m:

Table 6.33 The dimension a of the flute for the herring-bone gears

m (mm)	a (mm)
2	32
2.5	38
3	42
3.5	48
4	53
5	60
6	67
7	75
8	85
10	100

The dimensions (mm) of other structural components of herring-bone gears are

$$l_{hu} = b_2 + a\,;$$
$$C = (0.3\text{–}0.35)(b_2 + a)\,;$$
$$S = 2.2m + 0.05(b_2 + a)\,;$$
$$h = 2.5m\,;$$
$$S_b \approx 6m\,;$$
$$S_0 = (1.0\text{–}1.1)S_b\,;$$
$$t = 0.35S_b \geq 3\,\mathrm{mm}\,.$$

To decrease consumption of high-quality steel, wheels are sometimes produced to be stackable. Gear rings from alloy steel are pressed on a center made of carbon structural steel (Fig. 6.70b). The structure of a composite wheel with two gear rings is shown in Fig. 6.70c.

6.6.2 Spur Gears with Internal Toothing

The dimensions d_{hu}, l_{hu}, S, and f of the main structural components (Fig. 6.71) of a wheel with internal toothing are taken according to the ratios for wheels with external toothing. The embodiment of wheels with internal toothing can be achieved according to one of the choices shown in Fig. 6.71a,b. They differ in the hub position relative to the gear ring: in Fig. 6.71a the hub is positioned inside the wheel, which provides better operating conditions of the toothing as compared with the choice illustrated in Fig. 6.71b, in which the hub is outside the gear ring. The version shown in Fig. 6.71a can be used when the distance from the outer surface of the hub to the internal surface of the gear ring is greater than the external diameter D_e of the shaping cutter with which the teeth are produced. Moreover, it is required that the pinion which will mesh with the wheel can be easily positioned between the gear ring and the hub. The diameter D_e of the shaping cutter, the dimension a of the flute for the outlet of the shaping cutter and the chip disposability that form in tooth slotting, for the spurs take values depending on the module:

Table 6.34 The diameter D_e of the shaping cutter and the dimension a of the flute depending on the module m

m (mm)	D_e (mm)	a (mm)
1.5	54	
1.75	56	5
2.0	56	
2.25	54	
2.5	55	
2.75	55	6
3	60	
3.5	56	
4	112	7
5	110	
6	120	8
7	126	
8	128	9

The flute dimension a of helical wheels with internal toothing is increased by 30–40%. The depth of the flute in every case is assumed to be $h = 2.5\,m$, and the disk thickness C is $(0.3–0.35)b_2$.

6.6.3 Gear Clusters

Gear wheels with two, three, or four gear rings are applied in gearboxes of cars, tractors, and metal-cutting machine tools. These structures are called *gear clusters*. They are produced according to the form shown in Fig. 6.72a–d. Between the single gear rings flutes for the runout of the cutting tool (shaping cutter) are in-

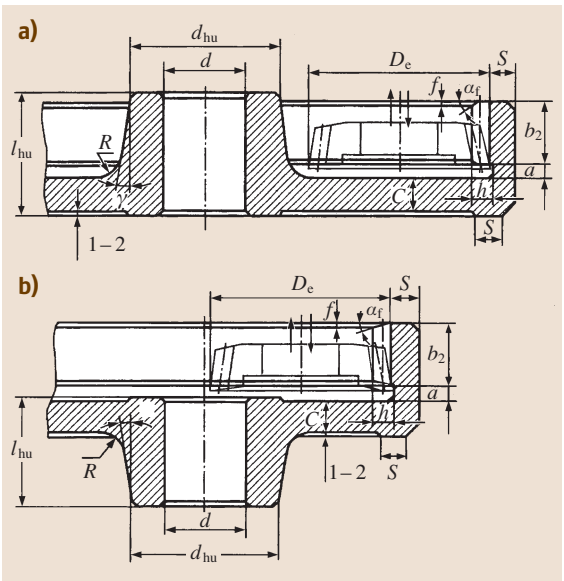


Fig. 6.71a,b Structural forms of the wheel of the internal toothing with hub position: **(a)** inside the wheel, **(b)** out of the contour of the gear ring

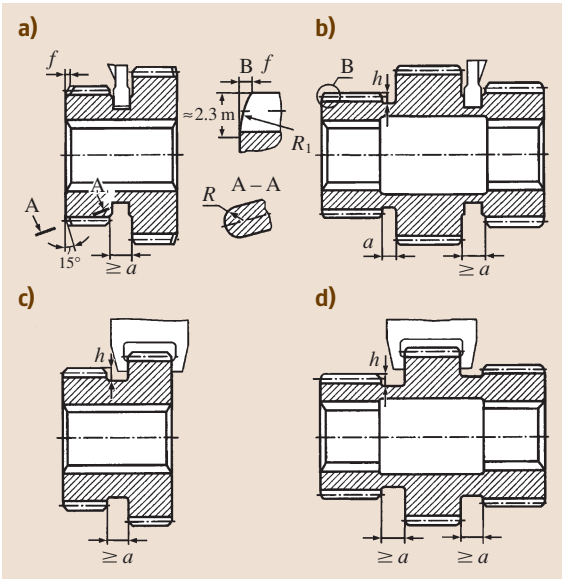


Fig. 6.72a–d Structural forms of one-piece gear clusters for gearboxes: **(a,b)** double, **(b,d)** triple

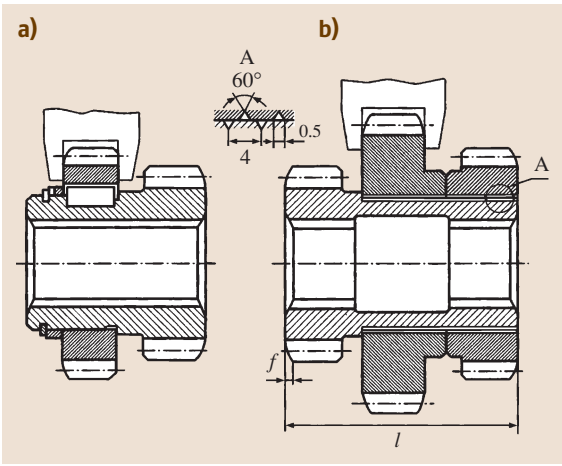


Fig. 6.73a,b Structural forms of composite gear clusters using (a) key joints and (b) adhesive joints

cluded. The flute width a is calculated depending on the diameter D_e of the shaping cutter (Sect. 6.6.2). The flute depth is $h = 2.5m$.

For the location of transferable lever or forks between the gear rings of the wheels circular grooves are made with a width that is more than a and with tolerance range H11, which is provided by grinding of the groove side walls in the case of wheel quenching. The teeth of the gear cluster rings from the entry side in the toothing are beveled $f = 0.6–0.7m$ on-the-miter $\approx 15°$ (Fig. 6.72a) and rounded (section A–A). Skewing is carried out along the curvilinear profile (remote element B, Fig. 6.72b). The gear ring teeth of mating gear wheels are also beveled and rounded from the entry side of the toothing.

The tooth work surfaces of high-accuracy gears (fifth and sixth accuracy degrees) are ground. For the exit of the grinding wheel a wide groove is needed, and therefore large axial dimensions of the gear clusters are used. To reduce these dimensions clusters are manufactured to be *stackable* through the use of key joints (Fig. 6.73a) or adhesive joints (Fig. 6.73b). In the latter case, helical grooves with mutually antithetical direction (remote element A) are cut on the mated surfaces for better toothing.

For a large cluster length ($l > 1.5d$) multispline (keyway) broaching is complicated, so the length of the opening is reduced with a recess in the middle part (Figs. 6.72b,d and 6.73b).

6.6.4 Bevel Wheels

The structural configurations of bevel wheels with an outer diameter of tooth tops of $d_{ae} \leq 120$ mm are shown in Fig. 6.74. With a pitch cone angle of $\delta \leq 30°$ the wheels are manufactured according to Fig. 6.74a. With an angle of $\delta \geq 45°$ they are manufactured according to Fig. 6.74b. If the pitch cone angle is 30–45°, both bevel wheel configurations are allowed. The hub dimensions d_{hu} and l_{hu} are determined from the ratios for cylindrical wheels. It is recommended to assume $l_{hu} = (1.2–1.4)d$.

Figure 6.75 shows bevel wheel configurations with an outer diameter of the tooth tops of $d_{ae} > 120$ mm.

For the approach illustrated in Fig. 6.75a the wheels are designed by *single-part* and *small-lot production*. Wheels with smaller diameters are produced from plain bars; wheels with greater diameters are manu-

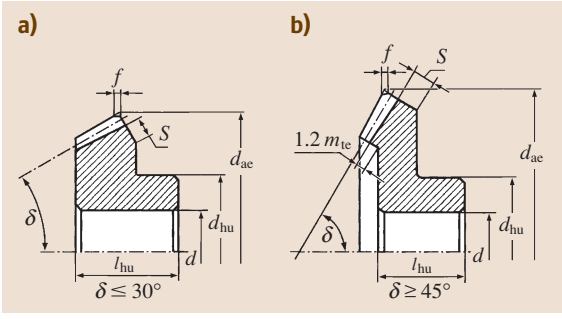


Fig. 6.74a,b Structural forms of conical gears with $d_{ae} \leq 120$ mm by: (a) $\delta \leq 30°$ and (b) $\delta \geq 45°$

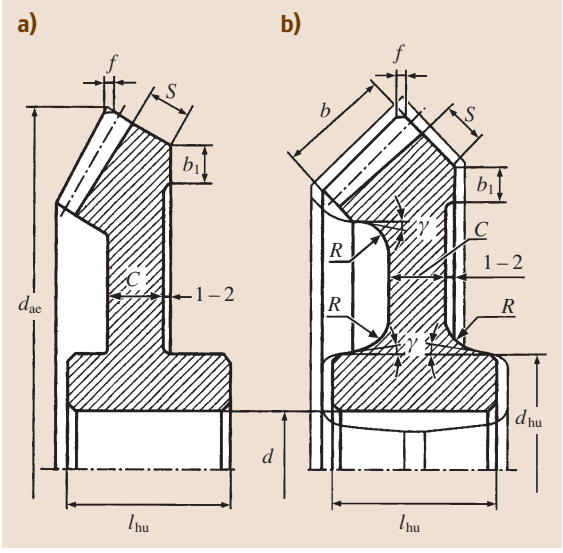


Fig. 6.75a,b Structural forms of conical gears with $d_{ae} > 120$ mm for (a) single part and (b) large lot

factured by open forging with subsequent turning. For the approach illustrated in Fig. 6.75b bevel wheels are designed for *high-volume applications*. The fine lines show a billet that results from forging in double-sided dies. The thickness C is determined from the ratios for cylindrical wheels.

With any wheel configuration the external tooth angles are blunted by a bevel of $f \approx 0.5 m_{te}$, when manufactured along the outer diameter d_{ae} parallel to the axis of the fit opening. The width S (mm) is assumed to be $S = 2.5 m_{te} + 2$ mm. A gear-ring end with width $b_1 \approx S$ is used for billet mounting for teeth cutting. Recesses with a depth of 1–2 mm are made to reduce the requirements for precise machining.

With an outer diameter of $d_{ae} > 180$ mm the wheels are sometimes manufactured to be stackable in order to save expensive steel. Depending on the wheel dimensions, the gear ring is fastened to the center by bolts, which are installed without clearance *for reaming* (Fig. 6.76a) or to the shaft flange by rivets (Fig. 6.76b). The gear ring is positioned so that the axial force acting in the toothing is directed towards the bearing flange. Centering of the gear ring is mostly carried out according to the diameter D (Fig. 6.76), but not D_0; here the centering accuracy is higher (for the same fit, the tolerance on the dimension D of the gear ring and the center, as well as the possible fit clearance, are lower). According to the technology used it is easier to obtain a precisely fitting ring opening, which is smooth, without a ledge, with a lower timetable for surface processing of a smaller diameter. Composite bevel wheels of the main gears in many cars have centering of the gear rings according to the diameter D.

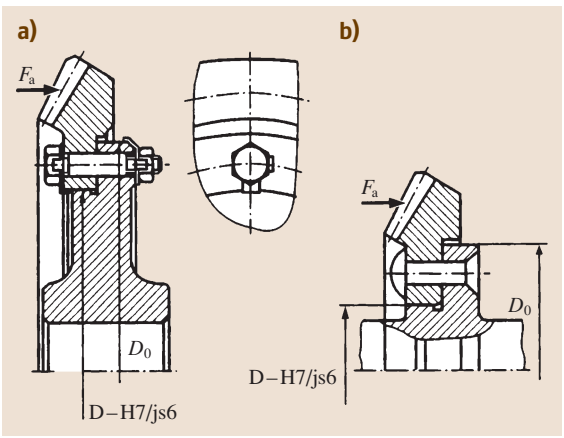


Fig. 6.76a,b Structural forms of composite conical gears using (a) a bolted connection and (b) a rivet joint

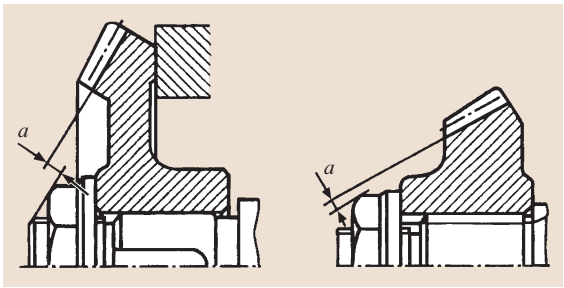


Fig. 6.77 Ensuring free outlet of the tool by cutting of circular teeth with a cutter head

However, centering according to D_0 results in higher joint rigidity, so as well as centering according to D such structures are applied where the centering of the gear ring is carried out in accordance with D_0.

Bevel wheels with circular teeth, which are cut with cutter heads by fastening the billet into the work holder, have wide application. It is necessary to include easy runout for the tool, with dimension of $a \geq 0.5 m_{te}$ (Fig. 6.77), where m_{te} is the external peripheral module.

6.6.5 Gear Shafts

There are two pinion embodiments of gearings: a single whole version with the shaft (*gear shaft*) and separately from it (*shell pinion*). The quality (rigidity, accuracy, etc.) of the gear shaft is higher, but production cost is lower than that of the shaft and shell pinion, so all the pinions of reduction gears are manufactured as a single whole with the shaft. Shell pinions are applied, for ex-

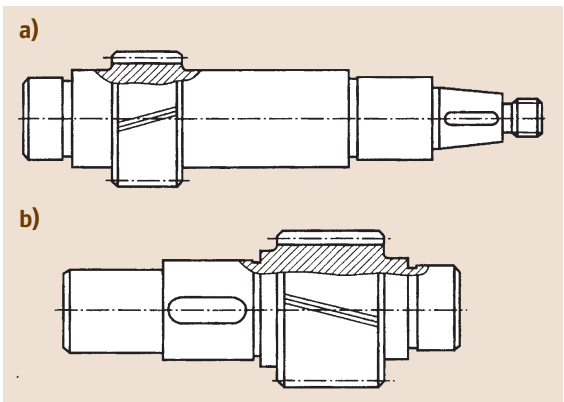


Fig. 6.78a,b Embodiment of the input (a) and idler (b) pinion shaft with free inlet and outlet of the gear-cutting tool

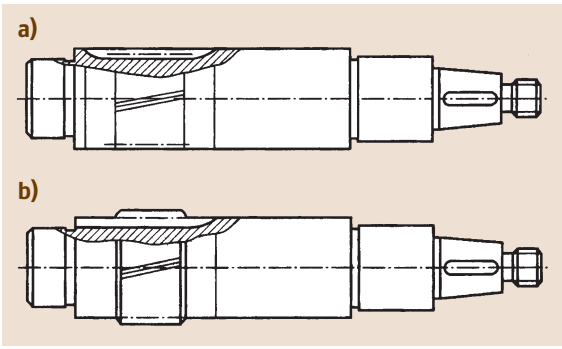


Fig. 6.79a,b Embodiment of the input shaft with a cutting-in pinion in the absence **(a)** and presence **(b)** of the free outlet of the tool from the side of the shaft collar

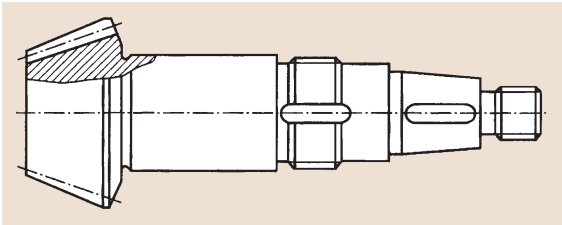


Fig. 6.80 Design version of the tapered pinion shaft

ample, in cases in which, under operating conditions, the pinion must be mobile along the shaft axis.

Figure 6.78 shows the structures of the gear shaft: (a) for high speed (with a small gear ratio) and (b) for low speed (countershaft) of the steps of the double-reduction gear. Both structures provide tooth cutting with easy exit of the tool.

With high gear ratios the outer pinion diameter, as a rule, differs only slightly from the shaft diameter, and the gear shafts are designed as illustrated in Fig. 6.79a,b. In this case, the teeth are cut on the surface of the shaft. The exit of the milling cutter is determined graphically according to its outer diameter D_f, assumed to be dependent on the module m:

Table 6.35 The outer diameter D_f on the milling cutter depending on the module m

m (mm)	D_f (mm) with accuracy degree	
	7	8–10
2–2.25	90	70
2.5–2.75	100	80
3–3.75	112	90
4–4.5	125	100
5–5.5	140	112
6–7	160	125

It is advisable to avoid cutting-in pinions, because in this case gear milling and tooth grinding are difficult. If possible, entry of the tool from the side of the shaft shoulder should be incorporated (Fig. 6.79b). A version of the tapered gear shaft is shown in Fig. 6.80.

6.6.6 Worm Wheels

Most often worm wheels are manufactured to be stackable: the center is made of gray iron or steel, the gear ring is made of bronze. The junction of the ring with the center must provide the transmission of high torque and relatively low axial force.

The design of the worm wheel and the technique used to bond the ring to the center depend on the numbers required. In single-part and small-lot production, when the annual numbers are less than 50 items, and with small wheel dimensions ($d_{ae2} < 300$ mm), the gear rings are joined to the center with an interference fit. For constant rotating sense of the worm wheel a collar is included where the axial force is directed (Fig. 6.81a). The junction of the gear ring with the center does not need a collar (Fig. 6.81b). In joints with relatively small interference the screws (usually three along the circle) are mounted in the butt of the gear ring and the center.

In the case of large wheel dimensions ($d_{ae2} \geq 300$ mm) mounting of the gear ring to the center can be carried out with bolts installed without clearance (Fig. 6.81c). In this case, the gear ring is previously centered according to the diameter D; mating of the centering surfaces is done according to the transition fit. Finally, the location of the gear ring is determined by mating of its opening with bolt bodies mounted without clearance. In this construction it is necessary to include reliable nut fastening; *retaining and spring-lock washers are not recommended*.

The rotational frequency of worm wheels is, as a rule, not high, and balancing is not carried out. Thus, inactive areas of the rim, disk, and hub of the wheel are left unprocessed and beveled with large rounded radii. Sharp edges on the ring ends are blunted with a bevel of $f \approx 0.5$ m with rounding to a standard value, where m is a module of the toothing. The dimensions of other structural components are

$$S \approx 2m + 0.05b_2\,; \quad S_0 \approx 1.25S\,;$$
$$C = (1.2\text{–}1.3)S_0\,; \quad h \approx 0.15b_2; \quad t \approx 0.8h\,.$$

Other structural components of worm wheels should be assumed to be the same as for cylindrical wheels (Sect. 6.6.1).

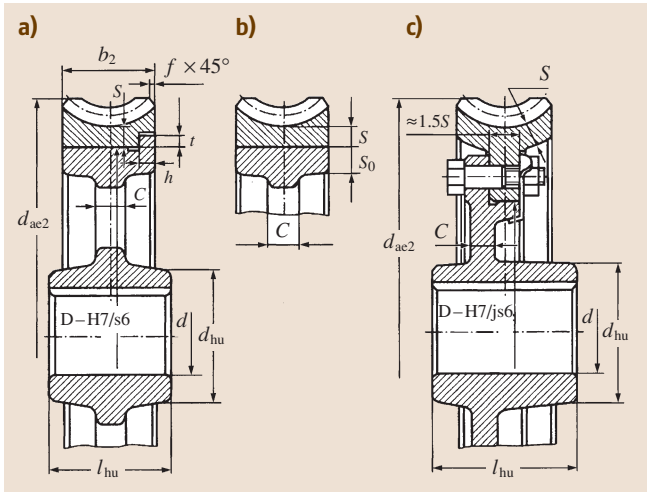


Fig. 6.81a–c Embodiments of the composite worm wheel using a joint with interference **(a,b)** and bolted connection **(c)** ▶

In the case of lot production (annual numbers of more than 100 items) it is economically sound to apply a build-up ring, as the working accuracy requirements of the mated surfaces of the ring and the center reduce, presses are not needed for joining, and screw fixing is not required.

An iron or steel center heated to 700–800 °C is put into the die, warmed to 150–200 °C and coated with molten bronze. Upon cooling there is interference between the center and the ring, caused by shrinkage of the freezing of the molten metal of the ring. Six to eight hollows of different forms are included on the rim of the center; after building-up, lugs form that additionally transfer both the peripheral force as well as the axial force. The thickness of the build-up ring is assumed to be $S \approx 2m$. The outer surface of the center is obtained by cutting or chill casting.

Figure 6.82 shows configurations of worm wheels where the centers are obtained by cutting. A concave surface of the center (Fig. 6.82a,b) is obtained by turning. The difference between these two versions lies in the form of the transverse grooves obtained by means of radial infeed of the milling cutter: (a) a disk cutter (the rotation axis of the cutter is perpendicular to the rotation axis of the wheel), and (b) a circular cutter (the rotation axis of the cutter is parallel to the rotation axis of the wheel). The dimensions of the grooves are $b \approx (0.3–0.5)b_2$ and $h = (0.3–0.4)b$. Both variants are equivalent in fabricability and labor output ratio. According to Fig. 6.82c the hollows on the rim of the center are drilled.

Figure 6.83 shows wheel configurations where the centers are obtained by chill casting. Machining of the outer surface is not carried out. Before the bronze is poured the center is cleaned of oil and oxide films by using chemical treatment. The variant shown in Fig. 6.83a, the chill mold design, is simpler, consisting of two parts only. In the variants shown in Fig. 6.83b and c the chill mold includes single segments, the number of which corresponds to the number of grooves. This elaborate

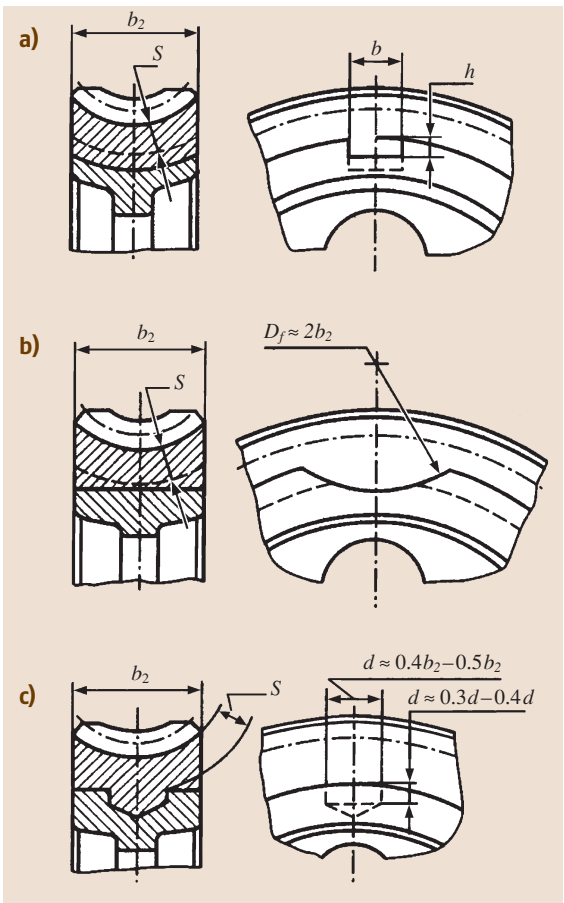


Fig. 6.82a–c Embodiments of the joints of surfaced gear rings with a center on whose rim there are valleys that are received through machining with **(a)** a disk cutter and **(b)** a circular cutter, **(c)** through drilling ▶

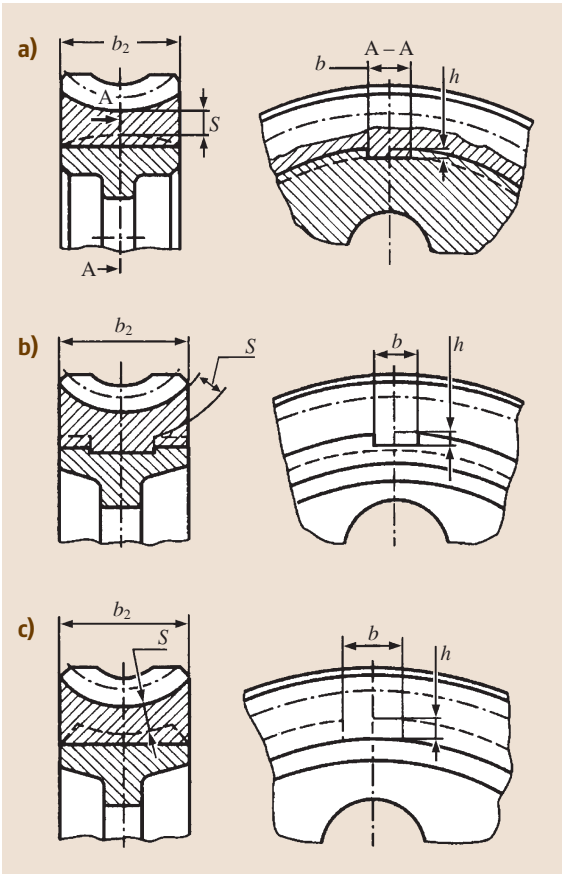


Fig. 6.83a–c Embodiments of the joints of the surface rings with a center received **(a)** by means of chill casting consisting of two parts segments **(b,c)** corresponding to the number of the grooves

design of the chill mold is due to the necessity to extract the billet after metal solidification. The dimensions b and h of the center grooves are assumed to be the same as in cutting.

The teeth of worm wheels have a concave form. This is why such a form of the outer center surface that follows the form of the teeth is optimal, as shown in Fig. 6.82a,b or in Fig. 6.83b,c. However, in practice other configurations are equally widely applied. With any configuration of the gear ring, machining and tooth cutting are carried out after joining of the ring with the center. The dimensions of other structural components are assumed to be given according to the relations given in Sect. 6.6.1

6.6.7 Worms

Worms are manufactured from steel, and mostly as a part of the shaft. The geometry of the worm including the length b_1 of the cut part and approximate distance l between the bearings is known from calculations and the layout diagram of reduction gears.

The dimensions of the salient from the reduction gear end of the worm shaft are adjusted using the appropriate dimensions of the motor shaft and joint sleeve. Then the shaft diameter is determined at the bearing seat.

Figure 6.84 shows possible structures of cylindrical worms. One of the main requirements is the provision of high worm rigidity. Toward this end the distance between the bearings is made as small as possible.

The diameter of the worm shaft in the uncut part is assigned to be of such kind to provide, as far as pos-

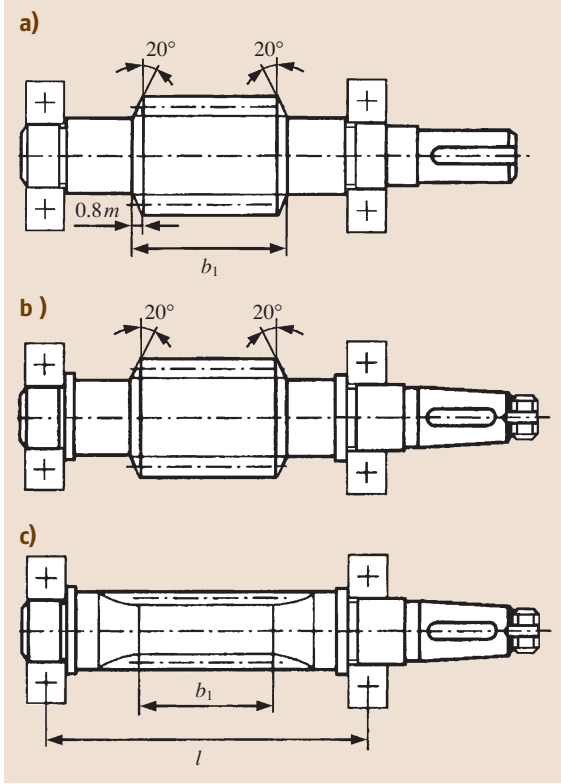


Fig. 6.84a–c Structural forms of cylindrical worms with a free inlet and outlet **(a,b)** of the tool by machining of the coils, **(c)** cutting-in design

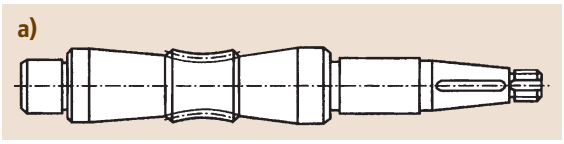


Fig. 6.85 Design version of the globoidal worm

sible, easy exit of the tool during coil processing and the necessary value of the thrust collar for the bearing. In Fig. 6.84a,b the diameter of the worm shaft before the cut part meets the condition of easy exit of the tool during coil processing. In Fig. 6.84a the collar height is both sufficient for the bearing thrust, and according to Fig. 6.84b, is low. Thus, for the bearing thrust a special collar is included.

Worms with a smaller diameter must be produced in accordance with Fig. 6.84c. In this case, the thrust collars are made at the bearing seats according to Fig. 6.84b as well as Fig. 6.84c.

Globoidal worms (Fig. 6.85) differ structurally from cylindrical ones in the form of the cutting area and the diameters of the bearing journals, which are comparable to the worm diameter. The other elements of this kind of worm are designed in the same way as for cylindrical worms.

6.6.8 Design Drawings of Gear and Worm Wheels: The Worm

The overall data for the design drawing development is given in Sect. 6.9.7. The design drawing shows an image component with the required dimensions, maximum dimensional deviations, form and position tolerances, roughness parameters, and specifications. In the right upper corner of the design drawings there is a table for the parameters of gear rings and worm coils. Figures 6.86–6.89 show drawings for cylindrical and bevel wheels, and worms and worm wheels, respectively.

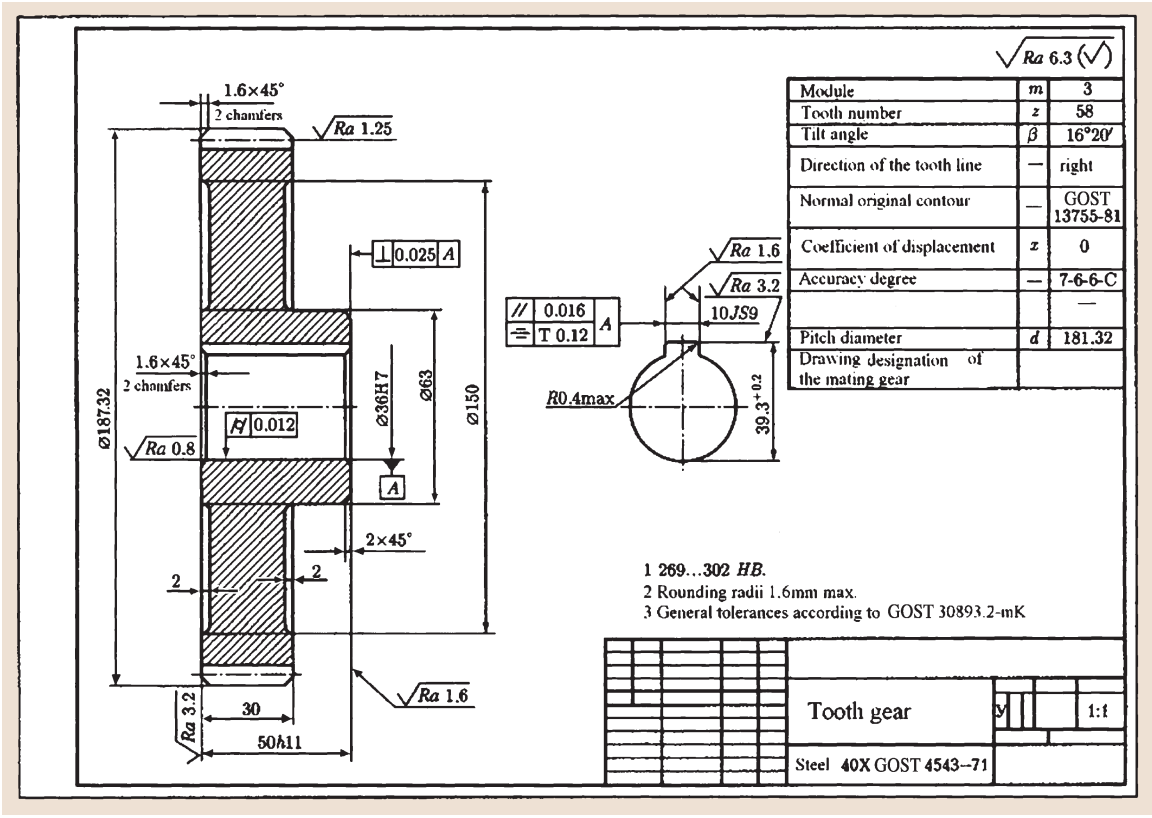


Fig. 6.86 Execution example of the working drawing of a spur gear

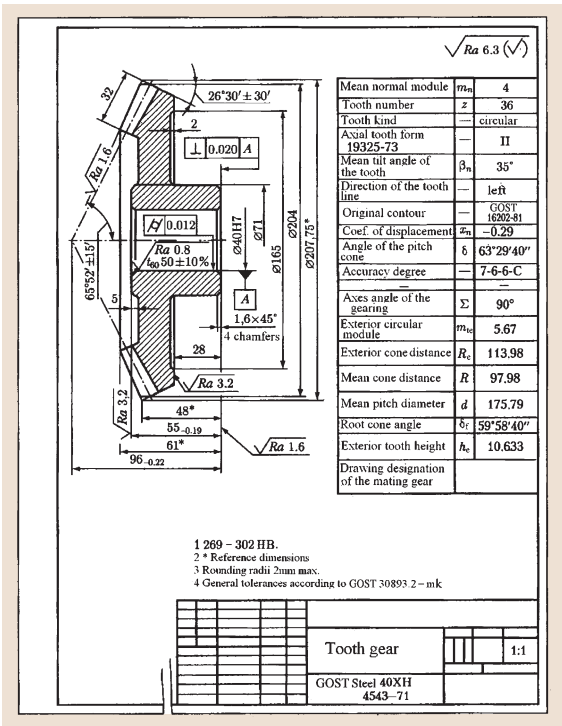


Fig. 6.87 Execution example of the working drawing of a conical gear

6.6.9 Lubrication of Tooth and Worm Gears

For lubrication of gearings the *oil bath system* is widely used. Oil is poured into the case of the reduction gear or the gearbox in such a way that the wheel rings are sunk into it. During rotation the wheels carry the oil and spray it inside the case. The oil falls on the inner walls of the case, from where it trickles down to the bottom again. A suspension of oil particles forms in the air inside the case and covers the surface of the details located inside the case.

Oil bath lubrication is applied when using a circumferential velocity of gear wheels and worms of 0.3–12.5 m/s. At higher velocities centrifugal forces throw oil off the teeth, and insufficient lubrication remains for toothing operation. Furthermore, under these conditions, the power power associated with oil stirring and temperature rises increases considerably. With circumferential velocities greater than 12.5 m/s circulating spray lubrication is used, where the toothing is lubricated by an oil jet created by a pipe nozzle under the pressure produced by a pump. The choice of lubricant is based on the operating experience of the corresponding machines.

Oils are mainly used for this application. The principle for the choice of the oil grade is as follows: the higher the circumferential velocity of the wheel, the lower the oil body; the higher the contact pressures in the toothing, the higher the body of the oil. Thus the required oil body is determined based on the contact stress and circumferential velocity of the wheels (Table 6.36). In Russia the oil grade for lubrication of tooth and worm gears is chosen according to Table 6.37.

In Russia the designation of machine oils consists of four groups of symbols, which indicates: (1) industrial (И), (2) according to purpose (Л, easily loaded units; Г, hydraulic systems; И, plain slideways; Т, heavily loaded units), (3) according to operating abilities (A, untreated oil; B, oil with antioxidant additives and stabilizers; C, oil with antioxidant and load-carrying additives and stabilizers; D, oil with antioxidant, load-carrying, and antiscoring additives and stabilizers; E, oil with antioxidant, load-carrying, antiscoring, and anti-"stick-slip" additives and stabilizers), and (4) the numerical kinematic viscosity class.

Table 6.36 Recommended kinematic oil viscosity for lubrication of toothed and worm gearings

Contact stresses σ_H (N/mm²)	Recommended kinematic viscosity (mm²/s) with circumferential velocity (m/s)		
	Up to 2	2–5	Over 5
For gearings at 40 °C			
Up to 600	34	28	22
600–1000	60	50	40
1000–1200	70	60	50
For worm gears at 100 °C			
Up to 200	25	20	15
200–250	32	25	18
250–300	40	30	23

Table 6.37 Kinematic viscosity of oils applied for lubrication of toothed and worm gearings

Oil grades	Kinematic viscosity (mm²/s)
For gearings at 40 °C	
И-Л-А-22	19–25
И-Г-А-32	29–35
И-Г-А-46	41–51
И-Г-А-68	61–75
И-ГТ-А-100	90–100
For worm gears at 100 °C	
И-Т-С-220	14
И-Т-С-320	20
Aircraft МС-20	20.5
Cylinder 52	52

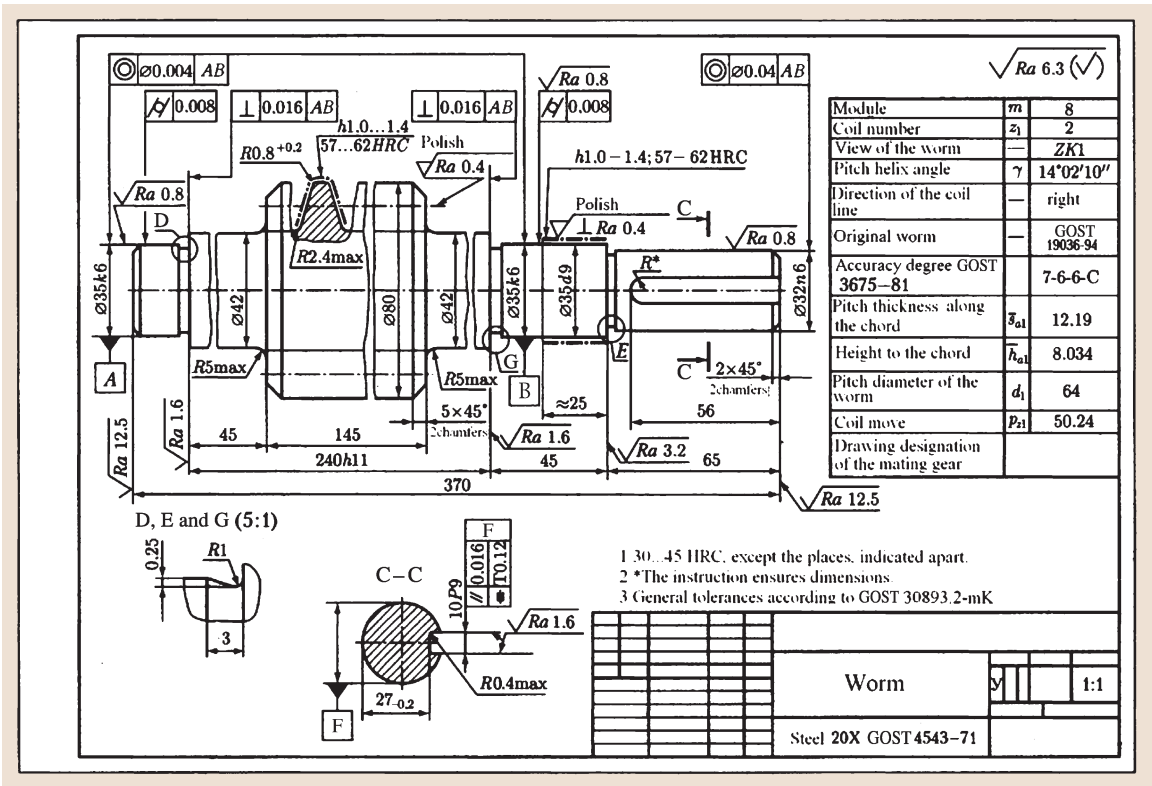


Fig. 6.88 Execution example of the working drawing of a cylindrical worm

6.7 Planetary Gears

6.7.1 Introduction

Planetary gears are gears that have gear wheels with moving axes. The most commonly used *ordinary single-row planetary gear* (Fig. 6.90) consists of a central wheel a with external teeth, a stationary central wheel b with internal teeth, planetary pinions g, wheels with external teeth interlocking simultaneously with a and b (here the planetary pinion number $n_w = 3$), and carriers h, to which the axes of the planetary pinions are fixed. The carrier is joined to the low-speed shaft. In planetary gearing *one wheel is locked* (connected to the case) [6.57–69].

With the stationary wheel b rotation of the wheel a causes rotation of the planetary pinion g relative to its own axis, and running of the planetary pinion along the wheel b moves the pinion axis and turns the carrier h. Thus the planetary pinion rotates relative to the carrier and rotates with the carrier around the central axis, i. e., it performs a motion similar to that of the planets, hence the name.

With wheel b stationary the movement is mostly transmitted from wheel a to the carrier h; transmission of the motion from the carrier h to the wheel a is possible.

Master links are links that are loaded with the external torque. For the gearing shown in Fig. 6.90 the master links are a, b, and h, i. e., two central wheels ($2K$) and a carrier (h). Such gearings are conventionally designated $2K$–h. The external torques on the links are denoted by T_a, T_b, and T_h. In planetary gears not only cylindrical wheels but also bevel wheels with a straight or oblique tooth can be applied.

If all the links in the planetary gear are mobile, i. e., both wheels and carrier, the gearing is called *differential*. With the help of the differential mechanism the

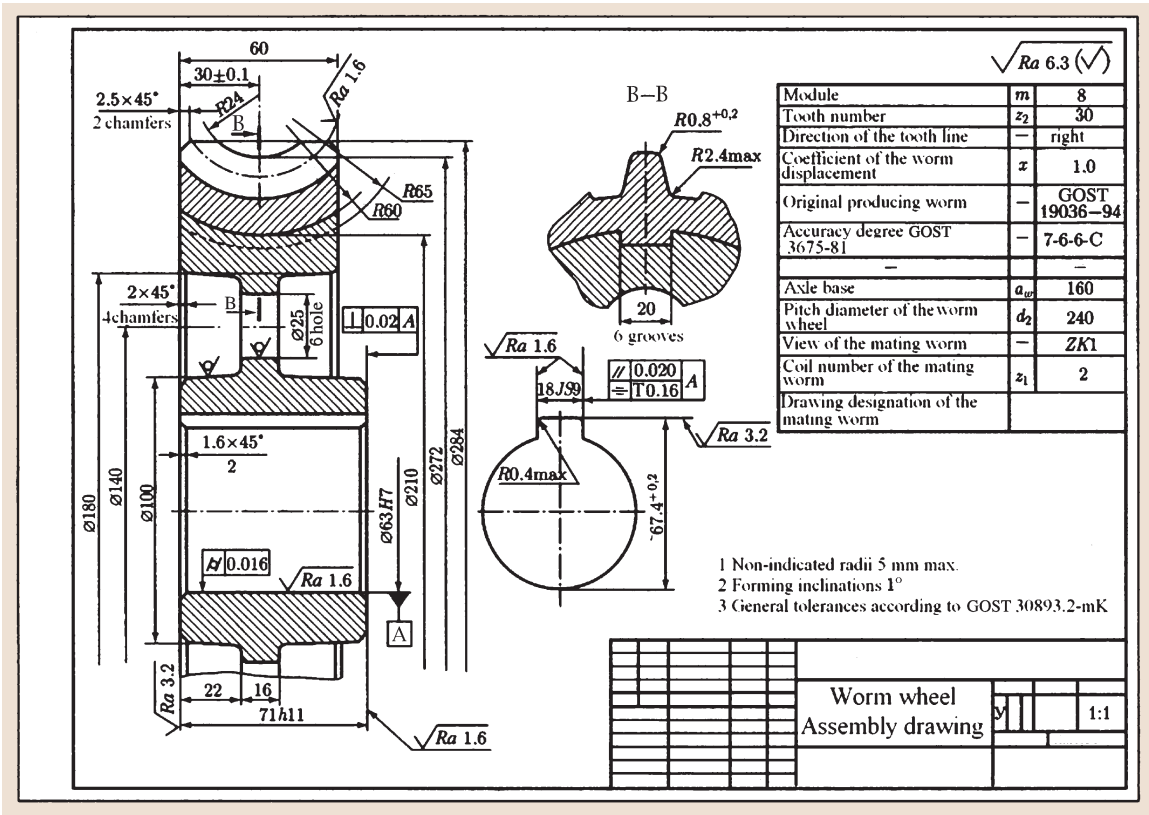


Fig. 6.89 Execution example of the working drawing of a worm wheel

motion of two links can be summed to one, or the motion of one link can be distributed between two links. For example, in the differential of the back axle of an automobile, the motion from the carrier h is transmitted simultaneously to wheels a and b, which allows one wheel to rotate more quickly than the other during turning.

The advantages of planetary gears are as follows:

1. Small dimensions and weight as a result of power transmission in several flows that are numerically equal to the number of planetary pinions. The load in each toothing is reduced by several times.
2. Convenience of assembly in machines due to the coaxiality of the drive shaft and the driven shaft.
3. Operation with less noise than standard gearings, due to the small wheel dimensions and force locking in the mechanism. In the case of symmetrical arrangement of the planetary pinions, the forces in the gear are mutually balanced.
4. Small loads on the shafts and bearings, which simplifies bearing construction and reduces losses.
5. Availability of high gear ratios with a small number of gear wheels and small dimensions.

The disadvantages are as follows:

1. Greater requirements of accuracy of manufacture and gearing mounting.
2. Greater number of details (bearings), with more complicated assembly.
3. To cut wheels with internal teeth, shaping cutters and gear shapers are needed, the base of which is smaller than for gear hobbers.

The planetary gear is applied as a reduction gear in power trains and devices, gearboxes where the gear ratio is changed by means of alternate braking of the different links (e.g., of the carrier or one of the wheels), and in differential gears in cars, tractors, machines, and devices. Planetary gears that are integrated with electric

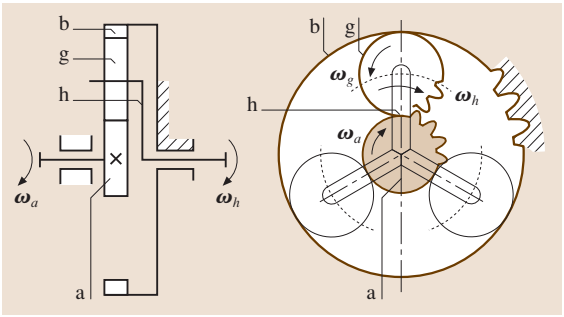


Fig. 6.90 Construction diagram of a one-row planetary gear

motors (e.g., reduction gear motors, hub motors) are often used.

6.7.2 Gear Ratio

For determination of the planetary gear transmission ratio the *method of the carrier stop* (Villis's method) is used. According to this method an additional rotation with angular carrier velocity ω_h is mentally imposed on the whole planetary gear, but in the reverse direction. Then the carrier acts as if stopped and the fixed wheel is released. There is also a so-called *reversed mechanism* representing a common nonplanetary gear, where the geometrical axes of all the wheels are stationary. Here the planetary pinions become intermediate wheels, which do not influence the transmission ratio of the mechanism. The transmission ratio in the reversed mechanism is determined as in case of a double-reduction gear with one external and one internal toothing.

The gear ratio relator has an essential value here. The transmission ratio u is assumed to be *positive* if the driven units and the drive units rotate in the same direction in the reversed mechanism, and *negative* if they rotate in different directions. Thus, for the reversed mechanism of the gear shown in Fig. 6.90 we have

$$u = u_1 u_2 = \left(-\frac{\omega_a}{\omega_g}\right)\left(\frac{-\omega_g}{-\omega_b}\right) = \left(-z_g/z_a\right)\left(z_b/z_g\right) = -z_b/z_a \,, \qquad (6.15)$$

where the tooth numbers of the appropriate wheels are indicated by z. In the reversed mechanism *the minus sign shows that wheels g and b rotate in the reverse direction with respect to wheel a*.

On the other hand, the imaginary stopping of the carrier in the motion transmission from a to b is equal to the subtraction of its angular velocity ω_h from the angular velocities of the wheels. Then for the reversed mechanism of this gear we have

$$u^h_{ab} = (\omega_a - \omega_h) \,/\, (\omega_b - \omega_h) = -z_b/z_a \,,$$

where $(\omega_a - \omega_h)$ and $(\omega_b - \omega_h)$ are, respectively, the angular velocities of the wheels a and b relative to the carrier h, and z_a and z_b are the tooth numbers of wheels a and b. The superscript "h" in the designation of the transmission ratio corresponds to the designation of the nonrotating unit, and the subscripts "a" and "b" correspond, respectively, to the drive element and the driven element. In this way the transmission ratio is determined from (6.15) for the planetary gear, which has a stationary carrier h ($\omega_h = 0$), a drive wheel a, and a driven wheel b. In the planetary gear any main element can be stopped.

For the planetary gear where the wheel b is fixed to be stationary in the case ($\omega_b = 0$), the wheel a is the drive, and the carrier h is driven. Thus from (6.15) we have

$$(\omega_a - \omega_h)/(0 - \omega_h) = -z_b/z_a \quad \text{or} \quad (-\omega_a/\omega_h) + 1 = -z_b/z_a \,.$$

It then follows that

$$u^b_{\mathrm{ha}} = \omega_a/\omega_h = 1 + z_b/z_a \,. \qquad (6.16)$$

For the planetary gear where the wheel b is fixed to be stationary in the case ($\omega_b = 0$), the carrier h is the drive, and the wheel a is driven. We then have

$$u^b_{ah} = \omega_h/\omega_a = 1/(\omega_a/\omega_h) = 1/(1 + z_b/z_a) \,.$$

Thus, depending on the stopped link, the planetary gear transmission ratio can take on different values, a characteristic of planetary gears that is used in gearboxes.

6.7.3 Planetary Gear Layouts

There are many different kinds of planetary gears. The *single-row gear* (Fig. 6.90) is most widely applied. This gear ($2K$–h) is physically simple and has small dimensions. It is applied in power and auxiliary drives. The efficiency factor of the gear is $\eta = 0.96$–0.98 with $u^b_{ah} = 3$–8.

To obtain higher transmission ratios *multiple* planetary gears are used in power drives. Figure 6.91a shows a planetary gearing consists of *two concatenated single-row planetary gears* with transmission ratios u_1 and u_2. The total transmission ratio u and the efficiency factor η in this case are

$$u = u_1 u_2 \le 64 \,; \quad \eta = \eta_1 \times \eta_2 = 0.92\text{–}0.96 \,.$$

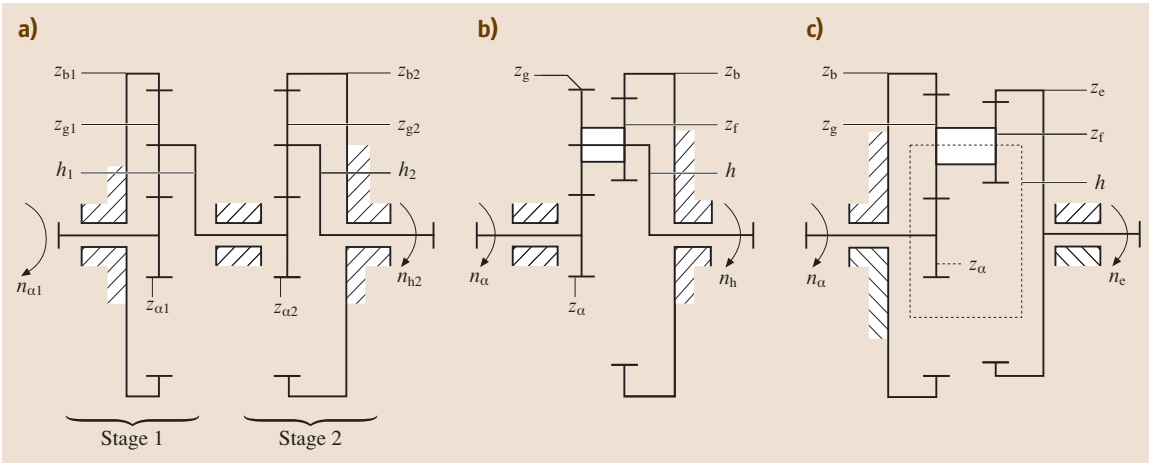


Fig. 6.91a–c Diagrams of planetary gears. (a) Double, (b,c) with double-row planetary pinion

Figure 6.91b shows the layout of a planetary gear with a double-row (two-ring) planetary pinion. The main units a and b are two central wheels, and the carrier h, and the gear $2K$–h. The transmission ratio for motion transmission from wheel a to carrier h and the fixed wheel b ($\omega_b = 0$) is

$$u^b_{ah} = \omega_a/\omega_h = n_a/n_h = 1 + z_b z_g/\left(z_f z_a\right) .$$

In this gear $u^b_{ah} = 3$–19 with $\eta = 0.95$–0.97.

In Fig. 6.91c the layout of the planetary gear $3K$ is shown. The main elements a, b, and e are three central wheels. The carrier h serves only to hold the planetary pinions. The transmission ratio for motion transmission from wheel a to wheel e and the fixed wheel b ($\omega_b = 0$) is

$$u^b_{ae} = \frac{\omega_a}{\omega_e} = \frac{n_a}{n_e} = \frac{(1 + z_b/z_a)}{\left[1 - z_b z_g/\left(z_f z_e\right)\right]} .$$

In this gear $u^b_{ae} = 16$–200 (up to 1600) with $\eta = 0.9$–0.7 (0.4).

6.7.4 Torques of the Main Units

For strength analysis of the toothing and force calculation in the toothing and for the following computation of the bearings it is necessary to know the torques on the main units. Further reasoning is provided for the planetary gear in Fig. 6.90. The torque T_a, (N m) on the main unit can be determined from the known power P_a (kW) and the rotational frequency $n_a(\mathrm{min}^{-1})$ according to

$$T_a = 9550 P_a/n_a .$$

From the balance condition we have

$$T_a + T_b + T_h = 0 .$$

From the condition of energy conservation with steady movement it follows that

$$T_a\omega_a + T_b\omega_b + T_h\omega_h = 0 .$$

With $\omega_b = 0$ we have

$$T_a\omega_a + T_h\omega_h = 0 \quad \text{and}$$
$$T_h = -T_a\omega_a/\omega_h = -T_a \times u^b_{ah} ,$$

or, taking into account losses η^b_{ah} due to movement of transmission from a to h,

$$T_h = -T_a u^b_{ah}\eta^b_{ah} .$$

The balance condition is written in the form

$$T_a + T_b + \left(-T_a u^b_{ah}\eta^b_{ah}\right) = 0 .$$

From this $T_b = T_a\,(u^b_{ah}\eta^b_{ah} - 1)$. In this way we obtain

$$T_b = -T_h\left[1 - 1/\left(u^b_{ah}\eta^b_{ah}\right)\right] .$$

Because u^b_{ah} is usually high, the torque on the central wheel b does not differ greatly from the torque on the output shaft: $T_b \approx -T_h$.

6.7.5 Toothing Forces

Peripheral forces in the toothing are determined from the torques (N m) and pitch diameters d_a, d_b, d_g (mm)

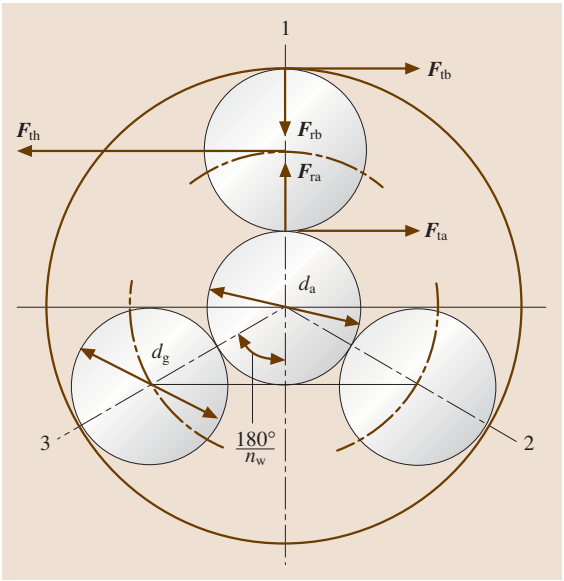


Fig. 6.92 Forces acting on one of the planetary pinions

of the gear wheels (for the gears without displacement)

$$F_{ta} = 2\times10^3 T_a k_w/(d_a n_w)\,,$$
$$F_{tb} = 2\times10^3 T_b k_w/(d_b n_w)\,,$$
$$F_{th} = 2\times10^3 T_h k_w/[(d_a + d_g) n_w]\,,$$

where n_w is the number of planetary pinions and k_w is a coefficient that takes into account the unevenness of the distribution of the torque between the planetary pinions (between the flows).

In Fig. 6.92 the forces acting on one of the planetary pinions are shown. The radial forces $F_{ra} = F_{ta}\tan\alpha_w$ and $F_{rb} = F_{tb}\tan\alpha_w$ are balanced here.

In the ideal mechanism $k_w = 1$ and the peripheral forces on wheel a in the toothing with all the planetary pinions equal: $F_{ta1} = F_{ta2} = F_{ta3}$. Wheel a is balanced (see the force polygon in Fig. 6.93a). However, in a real gear, as a result of production errors and detail deformation under loading, the torque is distributed irregularly between the planetary pinions, $F_{ta1} \neq F_{ta2} \neq F_{ta3}$, and the balancing force F_S acts on the shaft of the central wheel (Fig. 6.93b). Thus the values of the coefficient k_w are substantially higher than 1. To decrease the unevenness of the torque distribution and balance the peripheral forces wheel a is produced without bearings, i. e., it *floats*, and is connected to the drive shaft using a toothed coupling, which allows compensation for possible radial displacements Δ_r of the pinion (Fig. 6.93c). In this case, wheel a self-installs under the action of the force F_S, tending to reach equilibrium and thus overcoming the action of the frictional and inertial forces. Then the values of the coefficient k_w are much lower: $k_w = 1.05–1.15$.

6.7.6 Number Matching of Wheel Teeth

In contrast to traditional gearings the calculation of planetary gears starts with number matching of the wheel teeth. *The tooth number z_a of the central pinion a* is set based on the *non-undercutting condition* of the dedendum: $z_a \geq 12$. It is assumed that $z_a = 21–24$ for H $\leq$ 350 HB, that $z_a = 18–21$ for H $\leq$ 52 HRC; and that $z_a = 18$ for H $>$ 52 HRC. Tooth number matching of the other wheels is carried out taking into consideration three conditions: *coaxiality, mounting, and adjacency*.

Let us consider the tooth number matching by using the example of a spur planetary gear train without displacement (Figs. 6.90 and 6.91).

Layouts according to Figs. 6.90 and 6.91a

z_a is assumed to be in compliance with the above recommendations. *The tooth number z_b of the stationary central wheel b* is determined according to the *set*

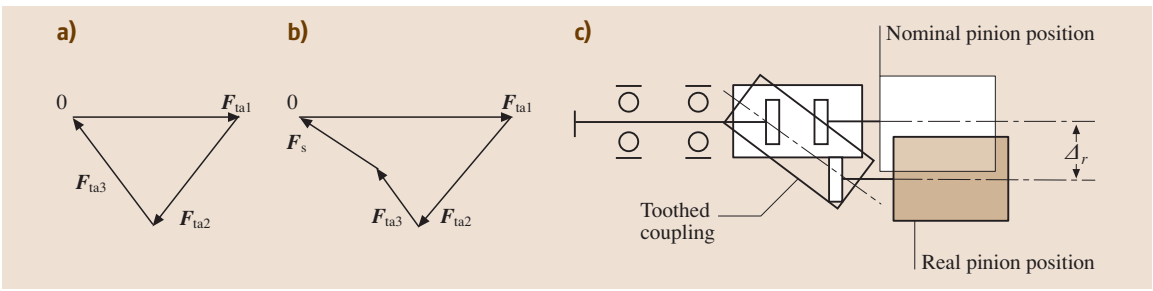


Fig. 6.93a–c Forces acting on the wheel a in the ideal (a) and the real (b) gear and (c) compensation of the radial displacement Δ_r with the help of the toothed coupling

transmission ratio u_{ah}^{b} from (6.16)

$$z_b = z_a \left(u_{ah}^{b} - 1\right) .$$

The tooth number z_g *of the planetary pinion* g is determined *according to the coaxiality condition*, in compliance with which the axle bases a_w of the gear sets with external and internal toothing are to be equal (Fig. 6.90)

$$a_w = 0.5(d_a + d_g) = 0.5(d_b - d_g) , \qquad (6.17)$$

where $d = mz$ is the pitch diameter of the appropriate gear wheel.

As the toothing modules of the planetary gear are equal, (6.17) takes the form

$$z_g = 0.5(z_b - z_a) .$$

Layout according to Fig. 6.91b

z_a is assumed. Then we have

$$z_b = z_a(u - 1)/c ,$$

where c is assumed to depend on the transmission ratio, according to:

Table 6.38 The value of factor c depending on the gear ratio u

u	c
10	1.4
12	1.5
14	1.6
16	1.8

The tooth number z_b after the calculation is rounded to a whole number that is divisible by the planetary pinion number. The coefficient is specified as

$$c = (u - 1)z_a/z_b .$$

Then we have

$$z_f = (z_b - z_a)/(c + 1) \quad \text{and} \quad z_g = cz_f .$$

For every layout the calculated tooth numbers are rounded to whole numbers. Furthermore, in accordance with Table 6.39, the coefficients of displacement x_1 of the pinion and x_2 of the wheel are chosen, and the coefficient B is determined as

$$B = 1000x_{sum}/(z_a + z_g) ,$$

where

$$x_{sum} = x_1 + x_2 .$$

According to the nomogram (Fig. 6.94) the toothing angle α_w of the gear is found.

Example

Determine the toothing angle with $z_a + z_g = 18 + 27 = 45$.

Solution

According to Table 6.39 we have $x_1 = 0.4$ and $x_2 = 1.02$, and consequently, $x_{sum} = x_1 + x_2 = 0.4 + 1.02 = 1.42$.

Then

$$B = 1000x_{sum}/(z_a + z_g)$$
$$= 1000 \times 1.42/(18 + 27) = 31.55 .$$

According to the nomogram (Fig. 6.94) we determine $\alpha_w = 26°55'$.

As the force calculation is not done and the modules of the toothing are unknown, for the layout in Fig. 6.91b

Table 6.39 Coefficients of displacement x_1 and x_2 of the pinions and the wheel in planetary gears

z_g	Values of the coefficients of displacements x_1 and x_2 with z_a											
	12		15		18		22		28		34	
	x_1	x_2	x_1	x_2	x_1	x_2	x_1	x_2	x_1	x_2	x_1	x_2
18	0.30	0.61	0.34	0.64	0.54	0.54	–	–	–	–	–	–
22	0.30	0.66	0.38	0.75	0.60	0.64	0.68	0.68	–	–	–	–
28	0.30	0.88	0.26	1.04	0.40	1.02	0.59	0.94	0.86	0.86	–	–
34	0.30	1.03	0.13	1.42	0.30	1.30	0.48	1.20	0.80	1.08	1.01	1.01
42	0.30	1.30	0.20	1.53	0.29	1.48	0.40	1.48	0.72	1.33	0.90	1.30
50	0.30	1.43	0.25	1.65	0.32	1.63	0.43	1.60	0.64	1.60	0.80	1.58
65	0.30	1.69	0.26	1.87	0.41	1.89	0.53	1.80	0.70	1.84	0.83	1.79
80	0.30	1.96	0.30	2.14	0.48	2.08	0.61	1.99	0.75	2.04	0.89	1.97
100	0.30	2.90	0.36	2.32	0.52	2.31	0.65	2.19	0.80	2.26	0.94	2.22
125	–	–	–	–	–	–	0.75	2.43	0.83	2.47	1.00	2.46

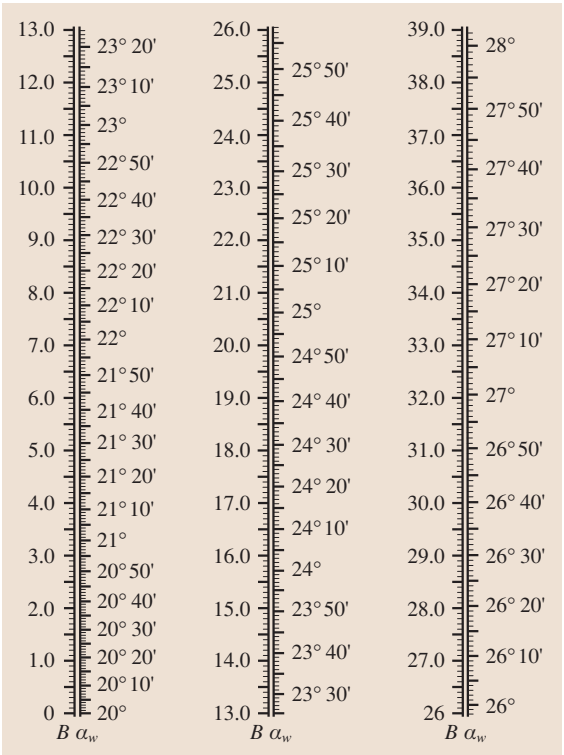


Fig. 6.94 Chart for the determination of the angle of action

the modules of both steps are assumed to be equal. Tooth numbers z_a, z_g, and z_b calculated in this way are checked according to the conditions of mounting and adjacency.

Layouts according to Fig. 6.90 and Fig. 6.91a

The Condition of Coaxiality

$$(z_a + z_g)\cos\alpha_{wa} = (z_b - z_g)/\cos\alpha_{wb}\ ,$$

where α_{wa} and α_{wb} are toothing angles of the gear with external (index a) and internal (index b) toothing. From this condition

$$z_g = \left(\frac{z_b}{\cos\alpha_{wb}} - \frac{z_a}{\cos\alpha_{wa}}\right)\frac{\cos\alpha_{wa}\cos\alpha_{wb}}{\cos\alpha_{wa} + \cos\alpha_{wb}}\ .$$

The mounting condition requires coincidence of the teeth with tooth slots to have place in all the toothings of the central wheels with planetary pinions, otherwise the gear cannot be mounted. It is determined that, with a symmetrical arrangement of the planetary pinions, the mounting condition is met when the tooth sum of the central wheels $(z_a + z_b)$ is divisible by the number of planetary pinions n_w (usually $n_w = 3$), i. e., $(z_a + z_b)/n_w = \gamma$, where γ is any whole number.

Layout according to Fig. 6.91b

The Coaxiality Condition

$$(z_a + z_g)/\cos\alpha_{wa} = (z_b - z_f)/\cos\alpha_{wb}\ .$$

Hence

$$z_f = (z_b/\cos\alpha_{wb} - z_a/\cos\alpha_{wa}) \\ /(c/\cos\alpha_{wa} + 1/\cos\alpha_{wb})\ ;$$

$$z_g = cz_f\ .$$

If, in the strength analysis of the gears according to the layout in Fig. 6.91b, different modules for the gears with external $(z_a - z_g)$ and internal $(z_f - z_b)$ toothing are assumed, the coaxiality condition for such a gearing is

$$(z_a + z_g)m_a/\cos\alpha_{wa} = (z_b - z_f)m_b/\cos\alpha_{wb}\ .$$

Hence

$$z_f = \frac{z_b m_b/\cos\alpha_{wb} - z_a m_a/\cos\alpha_{wa}}{c m_a/\cos\alpha_{wa} + m_b/\cos\alpha_{wb}}\ ,$$

where the tooth number $z_g = cz_f$.

Sometimes for fulfillment of the *coaxiality condition* it is convenient for one gear to be helical. The coaxiality condition in this case becomes

$$(z_a + z_g)m_a/(\cos\beta\cos\alpha_{wa}) \\ = (z_b - z_f)m_b/\cos\alpha_{wb}\ .$$

The required tilt angle β of the tooth is determined from this condition. *The mounting conditions* of the gear are then $z_a/n_w = \gamma$ and $z_b/n_w = \gamma$.

For all layouts of planetary gears control of the *adjacency condition* is carried out, which requires that the planetary pinions do not touch the teeth. To this end the sum of the top crest radii, which is $d_{ga} = m(z_g + 2)$, must be less than the distance l between their axes (Fig. 6.92), i. e.,

$$d_{ga} < l = 2a_w \sin(180°/n_w)\ , \qquad (6.18)$$

where a_w is an axle base.

For the layouts in Figs. 6.90 and 6.91a the axle base forms

$$a_w = 0.5m(z_a + z_g)\ ,$$

and in accordance with (6.18) the adjacency condition is fulfilled if

$$(z_g + 2) < (z_a + z_g)\sin(180°/n_w)\ .$$

The axle base of the gear, which is produced according to any layout, is

$$a_w = (z_a + z_g) m_a \cos\alpha / (2 \cos\beta \cos\alpha_{wa}) .$$

The actual values of the transmission ratios of reduction gears must not differ from the nominal values by more than 4% for single-reduction units, 5% for double-reduction units, and 6.3% for triple-reduction units.

6.7.7 Strength Analysis of Planetary Gears

The first calculation phases for planetary gears (choice of material and heat treatment, determination of allowable stresses) are performed in the same way as for traditional cylindrical gearings (Sect. 6.3.7).

Strength analysis is carried out for all toothings according to the formulas for traditional gearings. For example, for the gearing shown in Fig. 6.90 it is necessary to calculate the external toothing of the wheels a and g, and the internal toothing of the wheels g and b. The modules and forces of these toothings are equal and internal toothing is faster in accordance with its behavior, and therefore *when the same material is used for the wheels it is sufficient to calculate only the external toothing*.

Only the main characteristics of the calculation for planetary gears are examined below. For the determination of the allowable stresses $[\sigma]_H$ and $[\sigma]_F$ the service life ratios Z_N and Y_N are found according to the equivalent loading cycle numbers $N_{HE} = \mu_H N_k$ and $N_{FE} = \mu_F N_k$. *The number of stress cycles* N_k of the teeth for the whole lifetime is calculated *only for wheel rotation relatively to each other*.

For the *central pinion*

$$N_{ka} = 60 n_w n'_a L_h ,$$

where n_w is the number of planetary pinions, L_h is the total operating lifetime of the gearing (h), $n'_a = (n_a - n_h)$ is the relative rotational frequency of the drive central pinion, and n_a and n_h are the rotational frequencies of the central pinion and the carrier (min^{-1}).

According to n'_a the circumferential velocity is determined, in compliance with which the accuracy degree of the gear is set and the coefficients K_{HV}, K_{FV} are chosen.

For *planetary pinions*

$$N_{kg} = 60 n_t n'_g L_h ,$$

where n_t is the loading number of the tooth per revolution and $n'_g = n'_a z_a / z_g$ is the relative rotational frequency of the planetary pinion.

The tooth of the planetary pinion is loaded twice per revolution in the toothing with wheels a and b. However, by determination of the cycle number it is assumed that $n_t = 1$, because the *contact strength* analysis takes into account that the tooth of the planetary pinion works with wheels a and b with different flanks. By determination of *the allowable bending stresses* $[\sigma]_{Fg}$ for the teeth of the planetary pinion the coefficient Y_A is set, taking into consideration the double-sided application of the load (under a symmetrical loading cycle). The values Y_A are assumed to be $Y_A = 0.65$, 0.75, and 0.9, respectively, for refined, quenched with radiofrequency (RF) current heating (or cemented), and nitrided steels.

The axle base a_w of a spur planetary gear train for the wheel set of the external toothing (of the central pinion with the planetary pinion) is determined as

$$a_w = 450 \left(u' + 1\right) \sqrt[3]{\frac{K_H T_1 k_w}{\psi_{ba} u' [\sigma]_H^2 n_w}} ,$$

where $u' = z_g / z_a$ is a gear ratio of the calculated wheel set, $k_w = 1.05$–1.15 is an unbalance factor between the planetary pinions, $T_1 = T_a$ is the torque on the shaft of the drive central pinion (N m), n_w is the number of planetary pinions, ψ_{ba} is the coefficient of the face width of the wheel, with $\psi_{ba} = 0.4$ for wheel hardness $H \le 350$ HB, $\psi_{ba} = 0.315$ for $H \le 50$ HRC, and $\psi_{ba} = 0.25$ for $H > 50$ HRC.

The width b_b of the central wheel b is $b_b = \psi_{ba} a_w$. The width b_g of the planetary pinion ring is assumed to be 2–4 mm more than the value of b_b, and the width b_a of the central pinion is assumed to be $b_a = 1.1 b_g$. *The toothing module* is $m = 2a_w / (z_g + z_a)$. The calculated module is rounded to the nearest standard value and then the axle base is specified as $a_w = m(z_g + z_a)/2$. Bending analysis is performed according to (6.7) as for standard gearings.

6.7.8 Design of Planetary Gears

Figure 6.95 shows the construction of a single-reduction epicyclic unit produced according to the layout of Fig. 6.90. In this construction the central drive pinion is a floating link. In the radial direction the pinion self-installs along the planetary pinions. In the axial direction the pinion is fixed from one side with a pin butt (1) and from the other side with a toothed coupling (2) with spring rings (3) installed in it. The pitch diameter of the toothed coupling (2) is assumed, for simplicity of manufacture, to be equal to the diameter d_1 of the central pinion. The coupling diameter is $d_c \ge d_1 + 6$ m,

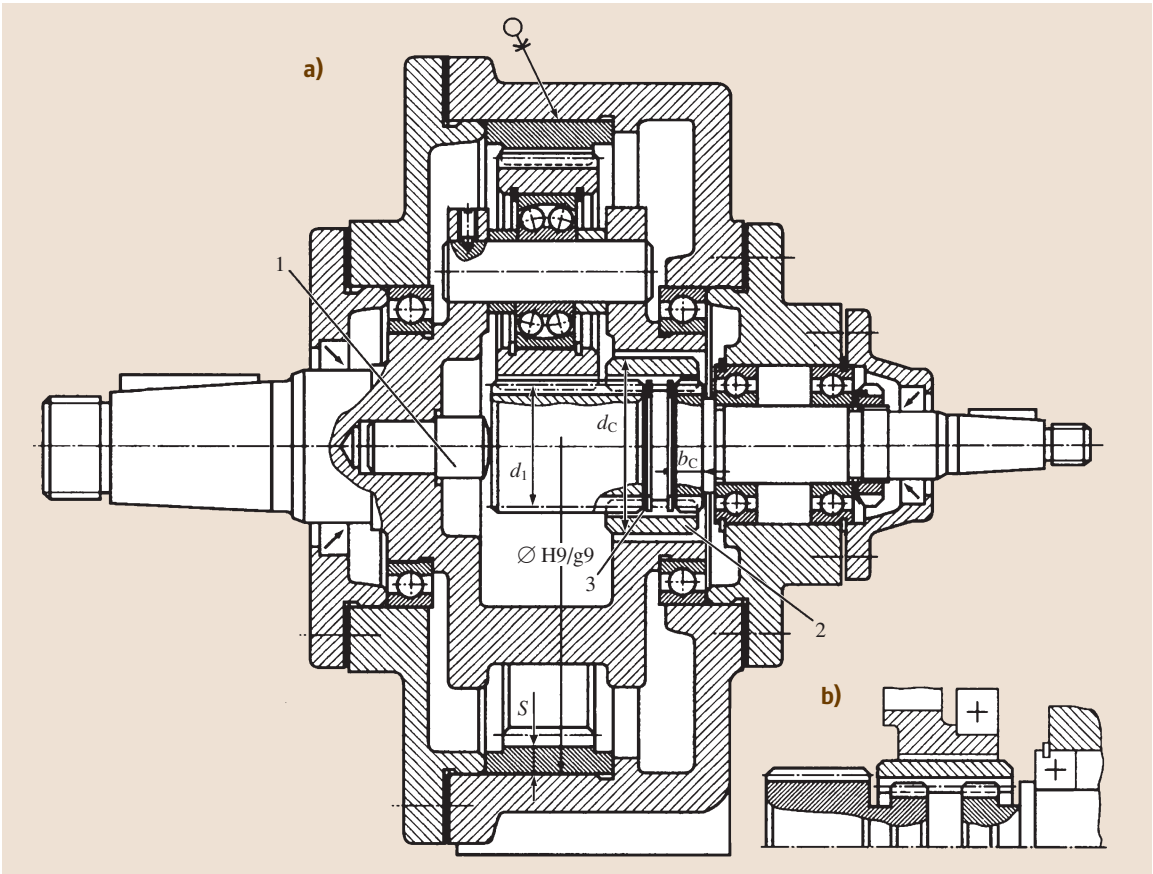


Fig. 6.95 (a) Planetary single-row gearbox according to the diagram of (b) execution version of the toothed coupling

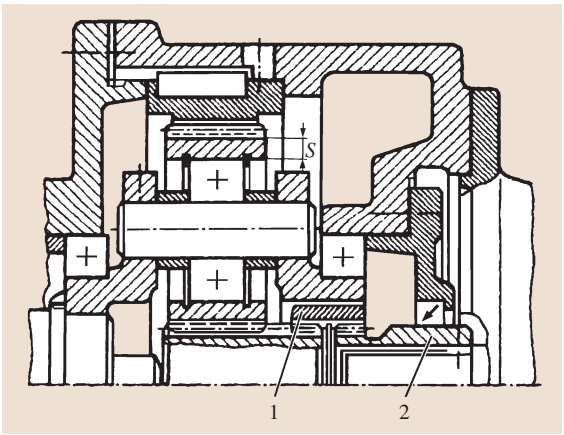


Fig. 6.96 Installation version of the toothed coupling on the shaft of the flange-mounted motor

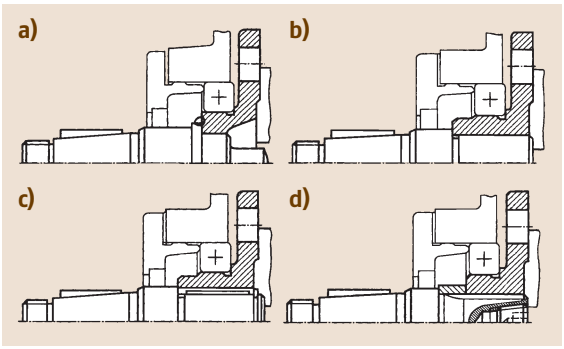


Fig. 6.97a–d Embodiments of the composite output shafts using (a) welded joints, (b) joints with interference, (c) key joints, and (d) spline joints

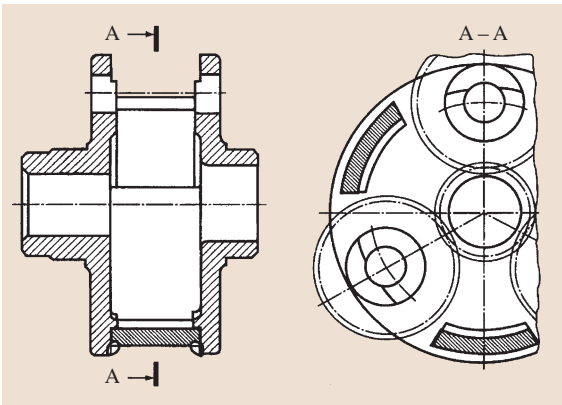


Fig. 6.98 Execution version of the welded carrier

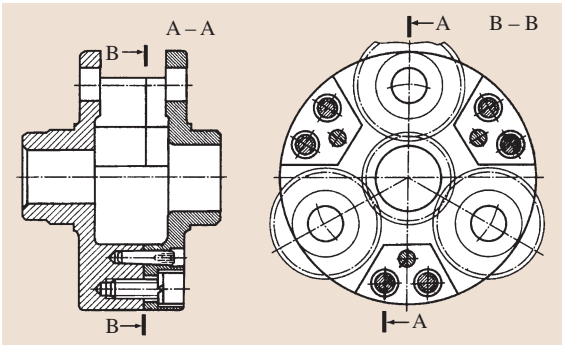
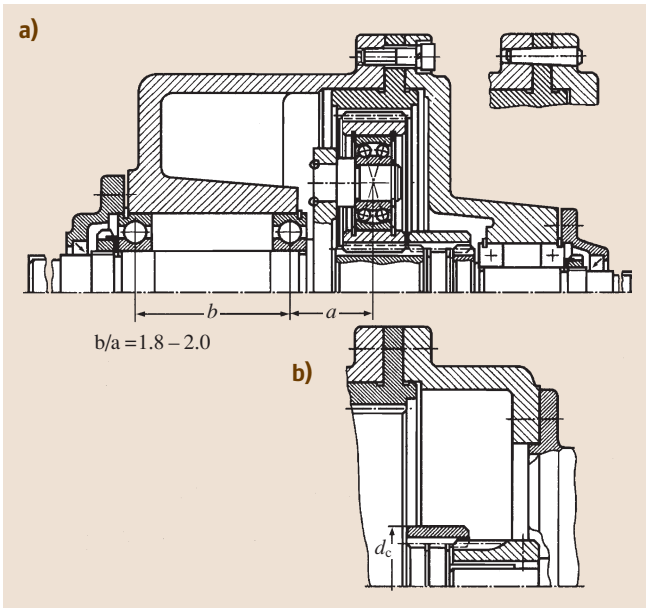


Fig. 6.99 Execution version of the composite carrier



the width of the toothing is $b_c = (0.2–0.3)d_1$, and the thickness of the stationary wheel is $S = 2.2m + 0.05b_b$, where b_b is the face width of the stationary wheel b. In cases when the coupling (2) is not built into the carrier opening, its outer diameter is reduced (Fig. 6.95b).

In Fig. 6.95 the input high-speed shaft is mounted on ball radial bearings with retaining snap rings. The input shaft receives movement from the electric motor through the *joint sleeve* mounted on the bevel or cylindrical shaft extension. By design of the reduction gear motor the toothed coupling (1) is connected with the pinion (2) installed on the shaft of the flange motor, as shown in Fig. 6.96.

For reduction of stress accumulation it is necessary for the planetary pinions to self-install along the stationary central wheel. To this end, radial spherical balls or roller bearings can be applied. The thickness S (Fig. 6.96) of the planetary pinion rim (mm) is $S \geq 2m + 1$, where m is the toothing module.

The low-speed shaft of the reduction gear is cast from high-duty cast iron (Appendix 6.A Table 6.95, cast irons) as a single unit with the carrier (Fig. 6.95) or in the case of small-lot or single-part production it is cast from steel and joined with the carrier through welding (Fig. 6.97a), interference fit (Fig. 6.97b), a key joint (Fig. 6.97c), or a spline connection (Fig. 6.97d).

Carriers are made as one-piece castings from steel or high-duty cast iron, as shown in Fig. 6.95, welded according to Fig. 6.98, or compositely fastened with six screws and three pins (Fig. 6.99).

In the structures indicated in Figs. 6.95, 6.98, and 6.99 the carriers are mounted in the case on two bearings, and the axis of the planetary pinion enters the openings of two walls of the carrier. Recently carriers are more often being designed with one wall, with the planetary pinion axles being cantilevered. Figure 6.100 shows the structure of an epicyclic unit with cantilever axles of the planetary pinions. In Fig. 6.100a the input shaft is connected to the electric motor shaft by means of a joint sleeve. In Fig. 6.100b the drive is obtained directly from the flange motor shaft. The carriers are mostly three-leafed (Fig. 6.101).

In this case, it is convenient to install the central pinion on the drive shaft using a spline connection or a serrated joint (Fig. 6.102). In order for this pinion to self-install the fits of the involute spline connection

Fig. 6.100a,b Planetary gearbox with cantilever position of the planetary pinion axles and **(a)** presence or **(b)** absence of the input shaft◀

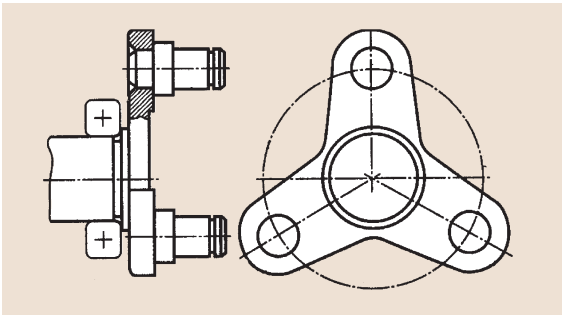


Fig. 6.101 Execution version of the three-leafed carrier

must have a large clearance along the centering surface (type H11/c11).

The wheel of the internal toothing withstands a considerable torque and must be tightly connected with the case. For this purpose the following are applied:

- Gluing of the wheel (Fig. 6.95) with an adhesive such as epoxy, phenol-formaldehyde, etc. The fit at a point of the wheel mating with the case is H9/g9.
- A key joint (Fig. 6.96).
- Flange fastening with screws and pins (Fig. 6.100).
- Mounting of three rolled or tapered pins along the circle (or in the radial direction) (Fig. 6.103); for air outlet *the flat* is taken off the roll pins (Appendix 6.A).

The simplest and modern solution is an adhesive joint.

Planetary gears made according to the layout in Fig. 6.91a differ from gears made according to the layout in Fig. 6.90 in the following two features:

1. The device for torque transmission from the carrier in the high-speed stage to the central pinion in the low-speed stage.
2. The case construction, in which many details are to be positioned, including two stationary wheels of the internal toothing.

All other elements of the epicyclic unit, e.g., the drive shaft, joint sleeves, planetary pinions, and carriers, are designed according to the same recommendations as for the elements of reduction gears of the layout in Fig. 6.90.

Torque transmission from the high-speed stage to the low-speed stage is carried out with the following methods:

- The spline shaft (1) produced as a unit with the central drive pinion of low-speed grade (Fig. 6.104).

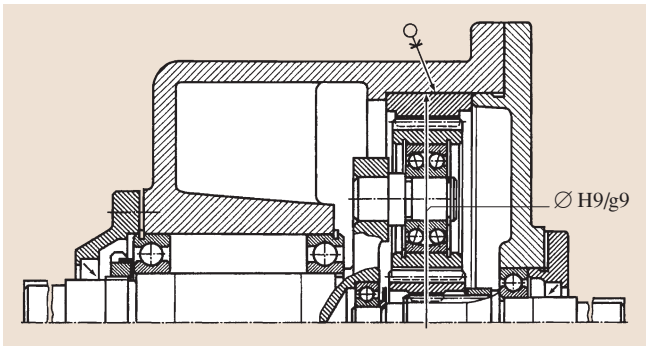


Fig. 6.102 Planetary gearbox with a central pinion mounted on the input shaft using a spline connection or a serrated joint

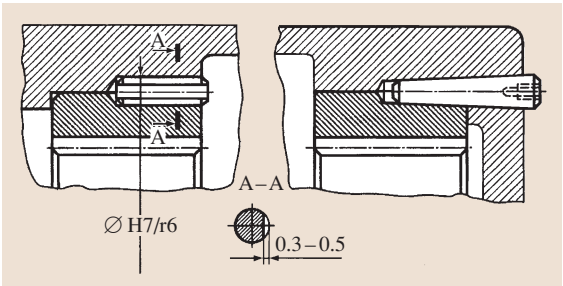


Fig. 6.103 Pin connection of the wheel of internal toothing to the case

- The toothed coupling (1) connecting the carrier of high-speed grade and the drive central pinion of low-speed grade (Fig. 6.105a,b).

In the middle of the reduction gear case a wall is foreseen, where the carrier bearings of high-speed and low-speed grades are positioned.

Planetary gears according to the layout in Fig. 6.91b differ from gears according to the layout in Fig. 6.91a

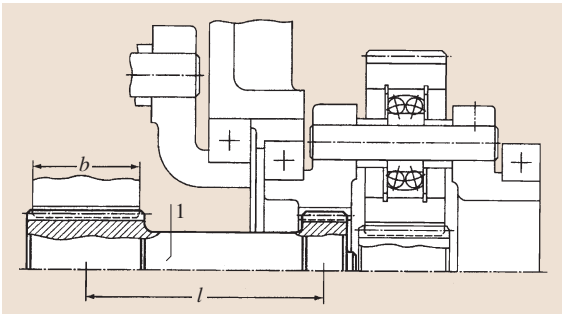


Fig. 6.104 Transmission of the torque from a high-speed step to a low-speed step with the help of a splined shaft

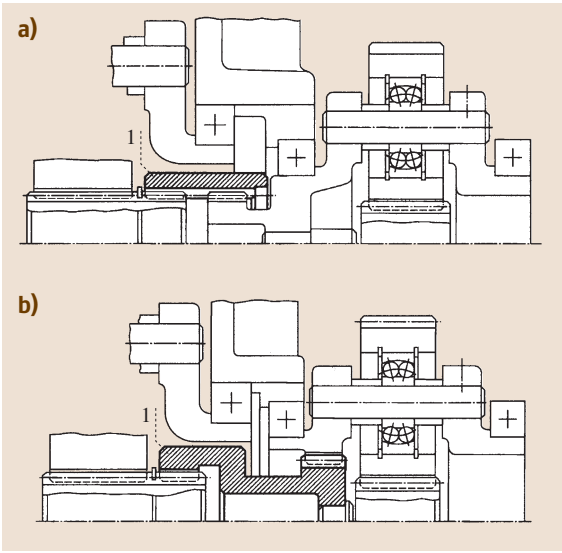


Fig. 6.105a,b Transmission of the torque from a high-speed step to a low-speed one with the help of coupling (a,b)

in the fact that the planetary pinions have two gear rings. Two bearings serve as supports of the planetary pinions; therefore the planetary pinions cannot self-install according to the central wheels. To decrease the stress accumulation along the tooth length the central drive pinion *a* is manufactured with barrel-shaped teeth, and the wheel *b* with the internal teeth is floating.

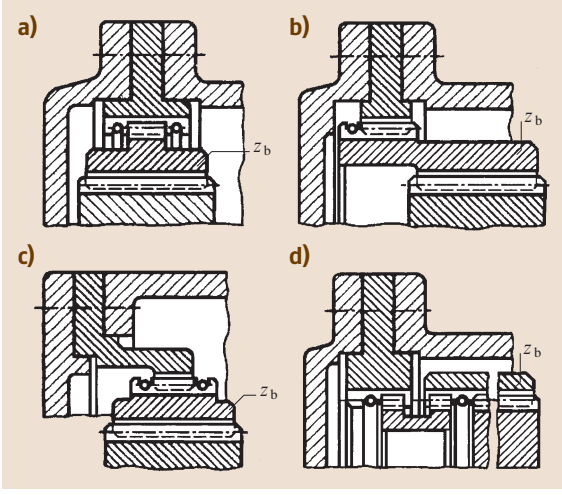


Fig. 6.106a–d Methods of connection of the floating wheel (a–d) with the case with internal toothing

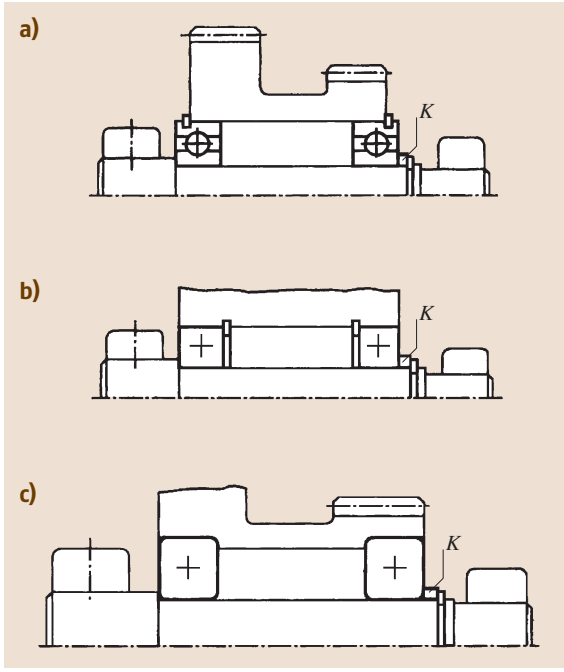


Fig. 6.107a–c Execution versions (a–c) of the supports of double-row planetary pinions

Depending on the location of the components of the planetary gear in the case the joining of the floating wheel *b* (tooth number z_b) with the other elements is carried out according to one of the choices shown in Fig. 6.106a–d. The other structural components of the planetary gears are designed according to the same recommendations as for gears having the layout shown in Fig. 6.90.

The performance variants of the planetary pinion supports are given in Fig. 6.107. The simplest solution is given in Fig. 6.107a. Instead of ball radial bearings, radial double-row spherical ball or roller bearings can be applied (Fig. 6.107b,c). In the supports of the planetary pinions, tapered roller bearings are also used, although rarely, because unit dismantling is needed for their adjustment. If it is not possible to insert the bearings given in Fig. 6.107 into the planetary pinions, needle (Fig. 6.108a) or friction bearings (Fig. 6.108b) are used.

In some epicyclic units constructions of planetary pinions with rotating axes are used. In Fig. 6.109a the simplest design is shown. In manufacture according to Fig. 6.109b radial double-row spherical ball or roller bearings can be applied as supports. Radial bearings with short cylindrical rollers are also used. The axis can be smooth, of constant diameter.

In all the versions shown in Fig. 6.109 matching or grinding of the compensatory rings K provides accuracy of the axial position of the details.

In order that the planetary pinions do not rotate relative to the axis, they are installed on the axis with slight interference (Fig. 6.110a), held with a set screw (Fig. 6.110b), or a roll pin (Fig. 6.110c).

Matching of Rolling Bearings

For support of the central shafts, ball radial single-row bearings are usually assumed, whereas for supports of planetary pinions, ball and roller spherical bearings are usually assumed. For calculation of the rolling bearings the support reactions F_{r1} and F_{r2} are defined. The main analytical models are given in Fig. 6.111.

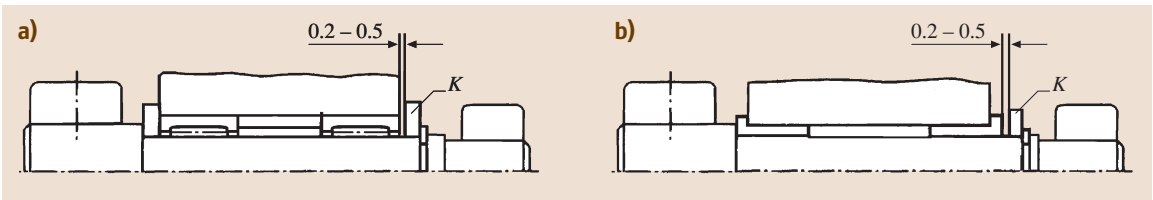


Fig. 6.108a,b Execution versions of the supports of double-row planetary pinions using (a) needle or (b) friction bearings

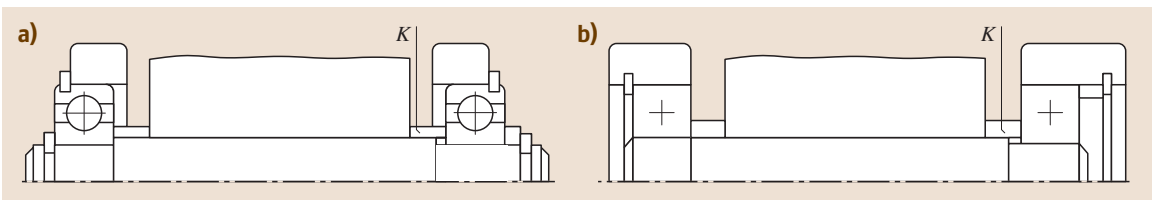


Fig. 6.109a,b Execution versions (a,b) of the supports of the rotating axles of planetary pinions

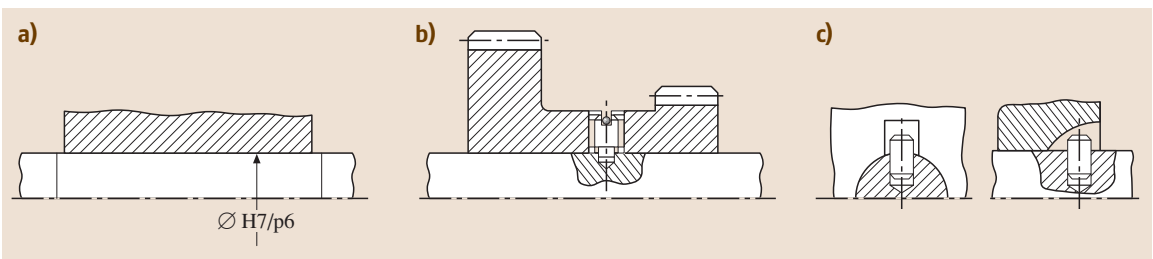


Fig. 6.110a–c Methods of fixation of planetary pinions on rotating axles. (a) By means of interference fit, (b) set screw, and (c) roll pin

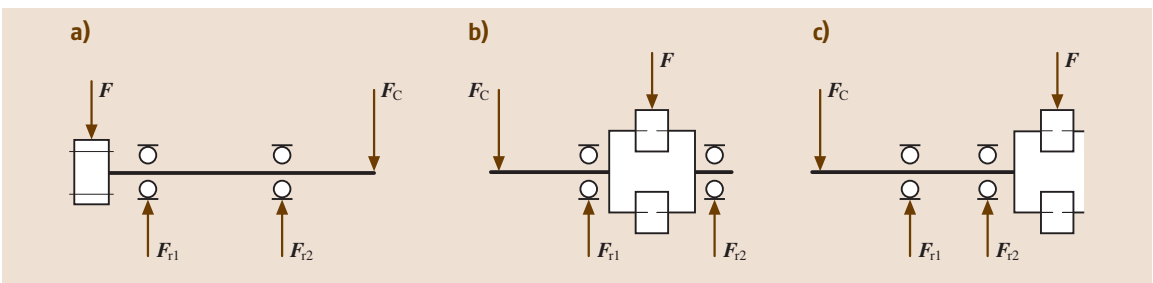


Fig. 6.111a–c Analytical models for matching of the bearings of the (a) input and (b,c) output shafts

Table 6.40 Recommended values of the coefficient a_{23}

Bearings	Values of the coefficient a_{23} for use conditions		
	1	2	3
Ball (except spherical)	0.7–0.8	1.0	1.2–1.4
Roller with cylindrical rollers, ball spherical double-row	0.5–0.6	0.8	1.0–1.2
Roller tapered	0.6–0.7	0.9	1.1–1.3
Roller spherical double-row	0.3–0.4	0.6	0.8–1.0

The input and output shafts of reduction gears are loaded with the force F acting from the side of the toothing and the cantilever force F_C (from the sleeve, belt drive, or chain gear). The assigned rolling bearings are calculated based on a set lifetime according to the action on the support reaction (F_{r1} or F_{r2}).

Taking into account the largest possible unevenness distribution of the total torque to the flows, the force F (N) acting on the shaft from the side of the toothing is determined from the following formulas:

For the input shaft (layout in Fig. 6.111a)

$$F = 0.2 \times 10^3 T_1 / d_1 ,$$

where T_1 is the torque on the shaft (N m) and d_1 is the pitch diameter of the toothed coupling teeth (2) (mm), which connects the input shaft to the drive pinion (Fig. 6.95).

For the output shaft (layouts in Fig. 6.111b,c and Figs. 6.95 and 6.100a)

$$F = 0.1 \times 10^3 T_h / a_w ,$$

where T_h is the torque on the output shaft (the carrier) (N m), $T_h = T_1 u \eta$, and a_w is the axle base of the gear (mm).

The bearings of the planetary pinions are the most heavily loaded

$$F_{r\,max} \approx 2 F_{t\,max} ,$$

where $F_{t\,max}$ is the peripheral force (N) and $F_{t\,max} = 2 \times 10^3 k_w T_{1\,max} / (n_w d_1)$. Here, $T_{1\,max} = T_1$ is the maximum of the long-acting (nominal) torque on the drive pinion (N m) and d_1 is the pitch diameter of the drive pinion (mm).

Table 6.41 Values of the coefficient a_1

Safety P_t (%)	Lifetime designation	Values of the coefficient a_1
90	L_{10a}	1
95	L_{5a}	0.62
96	L_{4a}	0.53
97	L_{3a}	0.44
98	L_{2a}	0.33
99	L_{1a}	0.21

The equivalent radial force for the bearing calculation under typical varying loading conditions is

$$F_r = K_E F_{r\,max} ,$$

where K_E is an equivalence coefficient.

The required radial dynamic load rating $C_{r,\,re}$ (N) of the planetary pinion bearings is determined from the formula

$$C_{r,\,re} = P_r \sqrt[k]{\frac{L'_{sah} n'_a z_a 60}{a_1 a_{23} 10^6 z_g}} ,$$

where $P_r = V F_r K_{dy} K_t$ is an equivalent radial load (N), $V = 1.2$ (the outer race rotates relative to the radial load), and L'_{sah} is the required bearing lifetime with given safety (h); $n'_a = (n_a - n_h)$ and z_a is the relative rotational frequency and tooth number of the central drive pinion, z_g is the tooth number of the planetary pinion, a_1 is a safety factor (Table 6.41), a_{23} is a use environment coefficient (Table 6.40: for spherical double-row ball-bearings $a_{23} = 0.5$–0.6, for spherical double-row roller bearings $a_{23} = 0.3$–0.4), $k = 3$ for ball bearings, and $k = 10/3$ for roller bearings.

6.8 Wave Gears

The wave gear is a power transmission in which rotation is transmitted by traversal of the deformation area of a flexible elastic section. A mechanical wave harmonic drive can be frictional and geared.

6.8.1 Arrangement and Operation Principles of Wave Gears

Frictional Wave Gears

The main elements of the gear are (Fig. 6.112a–c):

g – A flexible wheel representing a thin shell in the form of a cylinder with a bottom that is connected to the shaft

b – A rigid wheel connected to the case

h – A wave generator in the form of two rollers of major diameter, which are positioned on the carrier, connected with the high-speed shaft, and provide deformation of the flexible wheel

In practice there are also other embodiments of the main elements.

Figure 6.112a shows the relative position of the undistorted flexible g and rigid b wheels, with $d_g < d_b$ and $2W_0 = d_b - d_g$.

When installing the generator h the flexible wheel g is distorted, forming an ellipse (Fig. 6.112b). The power interaction of the wave gear details occurs at the contact points on the major ellipse axis. The generator presses the flexible wheel against the stiff wheel with sufficient force that the load is transmitted via frictional forces. With a stationary stiff wheel the rotation of the generator causes running-in of the flexible wheel to the stiff one and rotation of the flexible wheel in the direction opposite to the generator rotation.

Let us mark a point A on the flexible wheel g, at the contact with the rigid wheel b. After one revolution of the generator h clockwise, point A on the flexible wheel coincides with point A' on the stiff wheel (Fig. 6.112b), as the circumference of the flexible wheel with diameter d_g is less than that of the stiff wheel with diameter d_b. Therefore, the flexible wheel rotates in the direction opposite to the rotating sense of the generator. Figure 6.113 shows the flexible wheel before (circle) and after straining with the generator (ellipse). The maximum distortion in the direction of the major ellipse axis is indicated by W_0. The current distortion value W depends on the angle φ, which is measured clockwise from the major ellipse axis. The relation $W = f(\varphi)$ represents a wave function (Fig. 6.114). At the angle $\varphi = 2\pi$ two strain waves are observed and such a gear is called *two-wave*.

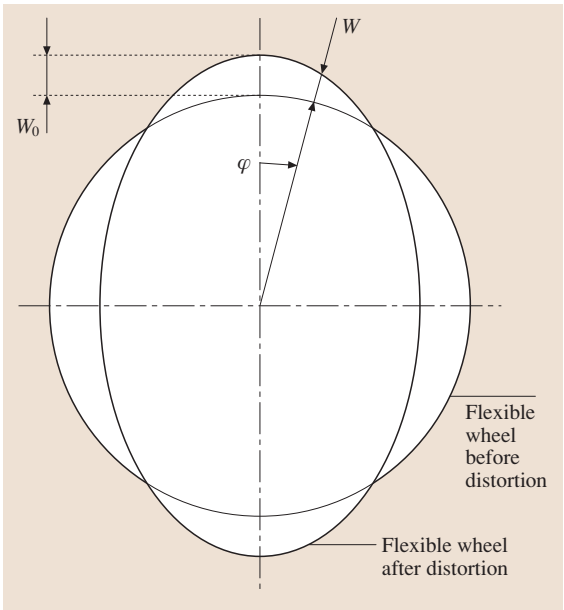


Fig. 6.113 Change of form of the flexible wheel due to straining

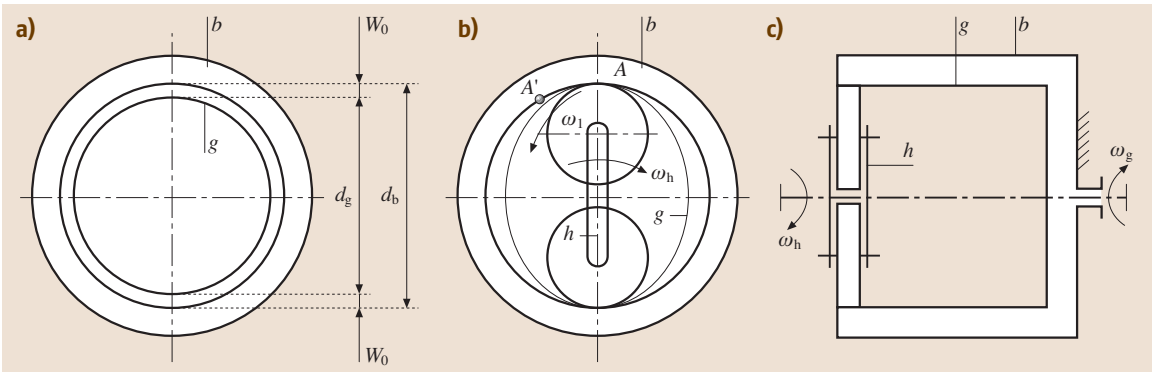


Fig. 6.112a–c Construction diagrams of the wave gear (**a**) before and (**b**,**c**) after straining of the flexible wheel

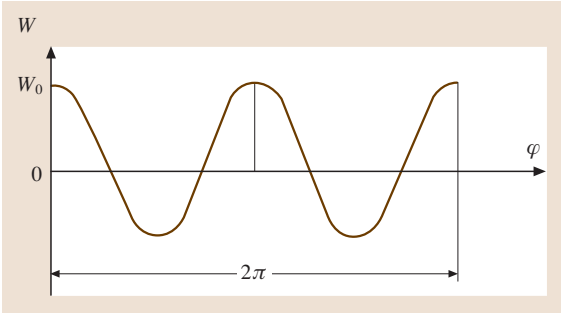


Fig. 6.114 Wave function of the deformation of the flexible wheel

Tooth Wave Gear

The flexible wheel g in such gears is a thin-walled cylinder. On its deformable end there is a gear ring with external teeth of involute profile. The flexible thin-walled cylinder plays the role of the resilient connection between the deformable O-ring seal and the stiff gear element, which can be the output shaft (Fig. 6.115a) or the case (Fig. 6.115b and c). *The rigid wheel b* has teeth with internal toothing. The tooth number z_b of the rigid wheel is greater than the tooth number z_g of the flexible wheel. *The wave generator h* represents a carrier (e.g., with two rollers) inserted into the flexible wheel. The flexible wheel, which distorts into the shape of an ellipse, forms two engagement zones along the major axis (Fig. 6.115b). In most cases, the generator is a pivotal element of the gear that is connected to the input shaft. The generator rotating with the angular velocity ω_h causes rotation of the flexible wheel with angular velocity ω_g (Fig. 6.115a) or of the rigid wheel with angular velocity ω_b (Fig. 6.115b,c).

In Fig. 6.116 the flexible wheel and tooth position before distortion are shown with hatches and without hatches after distortion. On straining of the flexible wheel with the generator under the action of the radial forces F_r the teeth of the flexible wheel move along the radius and engage with the teeth of the stiff wheel in the direction of the major ellipse axis. In the direction of the minor axis of the ellipse the deformation causes the teeth of the flexible wheel to move to the center and disengage from the teeth of the rigid wheel.

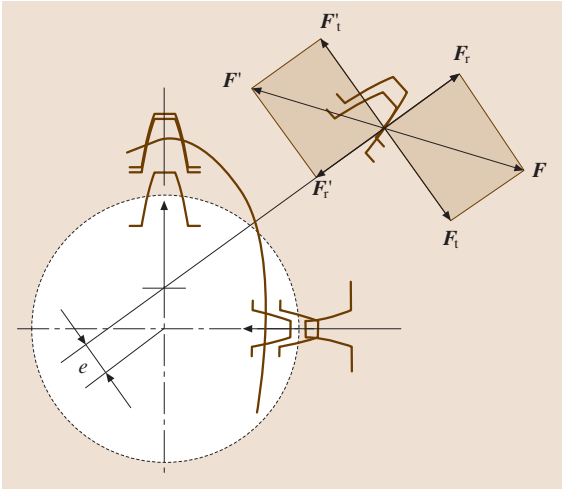


Fig. 6.116 Forces in the toothing of the wave gear

In the direction of the major axis of the ellipse meshing of the flexible wheel teeth with the rigid wheel teeth occurs along the entire depth of the tooth. By transfer from the major axis to the minor one the teeth of the flexible wheel gradually leave the mesh. A large number of the teeth are in simultaneous mesh: 25–40% of the teeth of the flexible wheel. In this way the gear can transmit a considerable torque even under slight load on every tooth.

With generator rotation by an angle φ a tooth of the flexible wheel moves in the radial direction by W, pressing the tooth of the rigid wheel with a force F that is directed along the normal line to the contact surfaces (Fig. 6.116). The force F can be represented in the form of the constituent forces: the peripheral force F_t and the

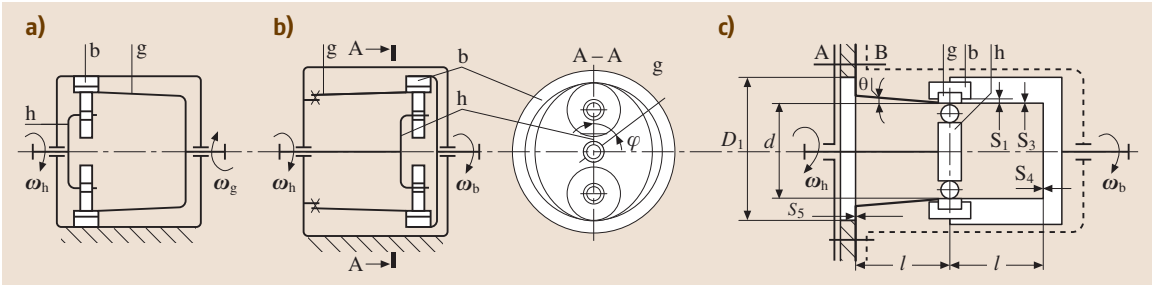


Fig. 6.115a–c Kinematic schemes of wave gears. **(a)** With a driven flexible wheel, **(b,c)** with a driven rigid wheel

radial force F_r. The reactions F_t' and F_r' act on the tooth of the flexible wheel.

If the generator is driving ($\omega_h \neq 0$), and the rigid wheel is fixed ($\omega_b = 0$), under the action of the force F_t' the flexible wheel rotates ($\omega_g \neq 0$) in the direction opposite to the generator rotation, as indicated by the minus sign in the formula for the transmission ratio, and as will be demonstrated hereinafter.

If the flexible wheel is stationary ($\omega_g = 0$), the rigid wheel rotates ($\omega_b \neq 0$) under the action of the force F_t in the direction of the generator rotation ($\omega_h \neq 0$), as indicated by a plus sign in the formula for the transmission ratio.

Figure 6.115c shows the layout of a wave gear with a stationary flexible wheel. Wave gears are the only power transmission that can transmit rotation *through the wall*, from a sealed space into a vacuum, without applying a rotatory seal. The flexible wheel g has the form of a blind sleeve with a flange, with which the wheel is fixed on the wall that separates the different media. The gear ring of the flexible wheel is in the middle sleeve part.

In wave gears each of the three primary elements can be driving. Thus, for example, in the case of a fixed flexible wheel and rotation of the rigid wheel in the counterclockwise direction, the flexible wheel acts upon the generator with a force F_r' (Fig. 6.116). The line of force action F_r' is along the normal to the curve that circumscribes the straining form of the flexible wheel. Under the action of the torque $T = 2F_r'\, e$ (where 2 is the number of strain waves) the generator rotates in the same direction as the rigid wheel.

6.8.2 Gear Ratio of Wave Gears

As with planetary gears, wave gears have three main elements that take external torques. Any main unit can be stopped.

1. The generator is stopped ($\omega_h = 0$). Rotation is transmitted from the flexible wheel with tooth number z_g to the rigid one (z_b), a common internal mesh

$$u_{gb}^h = \frac{\omega_g}{\omega_b} = \frac{z_b}{z_g} .$$

There is a plus sign in the formula, because the rotational directions ω_g and ω_b coincide.

2. The rigid wheel is stopped, $\omega_b = 0$ (Fig. 6.115a). This is the most frequent case (the standard wave gear).

Let us consider a differential wave gear with all three movable elements having angular velocities ω_g, ω_b, and ω_h. Let us choose a coordinate system that is quiescently bound to the generator. To do this assume that the angular velocity is $(-\omega_h)$ for the whole system. Then the elements have relative angular velocities

$$\omega_g - \omega_h\,; \quad \omega_b - \omega_h\,; \quad \omega_h - \omega_h = 0\,,$$

i. e., both wheels seem to rotate relative to the stationary generator. Then, as in the first case, we can write

$$u_{gb}^h = \frac{\omega_g - \omega_h}{\omega_b - \omega_h} = \frac{z_b}{z_g} .$$

If the rigid wheel is stopped, movement is transmitted from the generator to the flexible wheel and, therefore, $u_{hg}^b = \omega_h/\omega_g$ is determined. Supposing that $\omega_b = 0$ in the formula for the differential gear we have

$$\frac{\omega_g - \omega_h}{0 - \omega_h} = \frac{z_b}{z_g}\,;$$

$$-\frac{\omega_g}{\omega_h} + 1 = \frac{z_b}{z_g}\,;$$

$$u_{hg}^b = \frac{\omega_h}{\omega_g} = \frac{1}{(\omega_g/\omega_h)} = \frac{1}{1 - z_b/z_g}$$

$$= -\frac{z_g}{z_b - z_g} .$$

The minus sign shows that the sense of rotation of the flexible wheel is opposite to that of the generator.

3. The flexible wheel is stopped, $\omega_g = 0$ (Fig. 6.115b,c). Rotation is transmitted from the generator to the rigid wheel. It is necessary to find $u_{hb}^g = \omega_h/\omega_b$.

Supposing $\omega_g = 0$ in the formula for the differential gear we have

$$\frac{0 - \omega_h}{\omega_b - \omega_h} = \frac{z_b}{z_g}\,;$$

$$\frac{-\omega_h/\omega_b}{1 - \omega_h/\omega_b} = \frac{z_b}{z_g}\,;$$

$$-\frac{\omega_h}{\omega_b} = \frac{z_b}{z_g} - \frac{z_b\omega_h}{z_g\omega_b}\,;$$

$$\frac{\omega_h}{\omega_b} = \frac{-z_b/z_g}{(z_g - z_b)/z_g}\,;$$

Then

$$u_{hb}^g = \frac{\omega_h}{\omega_b} = \frac{z_b}{z_b - z_g} .$$

The senses of rotation of the generator and the rigid wheel coincide. The difference in the number of wheel

teeth is divisible by the wave number (as in planetary gears it is divisible by the planetary pinion number)

$$(z_b - z_g)/n_w = K_z \,,$$

where K_z is a whole number, with $u \geq 70 K_z = 1$, n_w is a wave number, and for the two-wave gear $n_w = 2$. Then

$$z_b - z_g = 2 \,.$$

Example

Determine the transmission ratios u^b_{hg} and u^g_{hb} for $z_g = 200$ and $z_b = 202$

$$u^b_{hg} = \frac{\omega_h}{\omega_g} = -\frac{z_g}{z_b - z_g} = -\frac{200}{202 - 200} = -100 \,,$$

$$u^g_{hb} = \frac{\omega_h}{\omega_b} = \frac{z_b}{z_b - z_g} = \frac{202}{202 - 200} = 101 \,.$$

6.8.3 Radial Deformation and the Transmission Ratio

From Fig. 6.112 it follows that $2W_0 = d_b - d_g$. For tooth wave gears with module m we have

$$2W_0 = d_b - d_g = m(z_b - z_g) \,.$$

Since $z_b - z_g = 2$, the radial deformation W_0 for wheels cut without displacement of the basic profile is $W_0 = m$.

For standard wave gears

$$u^b_{hg} = -\frac{z_g}{z_b - z_g} = -\frac{mz_g}{mz_b - mz_g} = -\frac{d_g}{d_b - d_g} = -\frac{d_g}{2W_0} \,.$$

In other words, *the transmission ratio in wave gears is equal to the ratio of the driven wheel radius to the difference of the radii of the rigid wheel and the flexible wheel or to the deformation dimension* W_0.

It follows that higher values of the transmission ratio u can be reached for low values of W_0, i. e., by small modules m. Major deformation dimensions W_0 correspond to lower values of u, for which the curvature of the flexible wheel and, therefore, bending stresses increase considerably in the toothing area.

The allowable range of the transmission ratio of the wave gear is

$$70 < u < 320 \,.$$

The lower limit on u is provided by the limit on the strength of the flexible wheel under bending stresses, whereas the upper limit is provided by the minimal module values ($m \geq 0.15$ mm).

Advantages of Wave Gears

1. The availability of a higher transmission ratio on one grade with a comparatively high value of the efficiency factor η. For one grade u up to 320 with $\eta = 0.7$–0.85.
2. The capability to transmit higher torques for smaller dimensions and mass due to the large number of teeth that engage simultaneously.
3. Operating smoothness and low kinematic inaccuracy due to two-zone and multipair toothing.
4. Rotation transmission from a sealed space without the use of rotatory seals.
5. Low loads on the shafts and bearings as a consequence of construction symmetry.
6. Operation with little noise.

Disadvantages

1. Production complexity of the thin-walled flexible wheel and the wave generator.
2. The need for special gear-shaping equipment to apply the small modules.
3. Limited rotational frequencies of the wave generator, leading to increased vibration.

Applications

Wave gears are applied in industrial robots and manipulators, in mechanisms with high transmission ratio, and also in devices with increased requirements of kinematic accuracy or tightness.

6.8.4 The Nature and Causes of Failure of Wave Gear Details

Some of the causes of failure of wave gear details are:

1. Fracture of the flexible wheel as a result of fatigue cracks in the tooth sockets, as the wheel is exposed to alternate bending stresses.
2. Bearing fracture of the wave generator as a consequence of the toothing force action and the resistance of the flexible wheel to deformation.
3. Skipping of the wave generator (rotation of the generator shaft without rotation of the output shaft) as a result of insufficient radial rigidity and great resilience of the wave generator and the stiff wheel for the transmission of high torques. Thereupon the teeth at the toothing entry rest with their tops against

each other, the stiff wheel bursts open, the generator is compressed, and skipping occurs.

4. Wear of the teeth. Insignificant wear of the teeth is caused by warping of the flexible wheel, which is deformable at one end; progressive wear is caused by tooth slipping through the entry into the toothing.

Alloy steel is used for *flexible wheels*. A billet in the form of a thick tube is refined with heat treatment. Strengthening with cold stiffening, including the tooth sockets, or with nitriding is recommended. Nitriding and cold stiffening of the gear rings are carried out after machining and tooth cutting. Cold stiffening increases the endurance limit by a factor of approximately 1.15, and nitriding increases it by a factor of about 1.4.

In Russia for *heavily loaded* flexible wheels (with low values of u) steels of heightened viscosity grades 38Х2МЮА (heat treatment, refinement, and nitriding, center stiffness 32–37 HRC, $\sigma_{-1} = 480–550\,\mathrm{N/mm^2}$), 40ХИМА (refining, 32–39 HRC, $\sigma_{-1} = 480–550\mathrm{N/mm^2}$), which are less sensitive to stress concentration, are used. *Intermediate-loaded and lightly loaded* flexible wheels are mostly produced from the steel grades 30ХГСИ (refined, 32–37 HRC, $\sigma_{-1} = 420–450\,\mathrm{N/mm^2}$, by further cloud-burst stiffening or nitriding $\sigma_{-1} = 480–500\,\mathrm{N/mm^2}$ (Appendix 6.A Table 6.95).

For welded flexible wheels steel grades such as 30ХГСА, 12Х18И10Т (18–22 HRC, $\sigma_{-1} = 280\,\mathrm{N/mm^2}$) are preferable) (Appendix 6.A Table 6.95).

The stiff wheels of wave gears are similar in construction to the wheels with internal teeth of standard and planetary gears. They are characterized by having a lower stress condition than flexible wheels. They are manufactured from standard structural steel grades 45, 40Х, and 30ХГСА with a stiffness that is 20–30 HB lower than that of flexible wheels. It is possible to produce stiff wheels from the iron grade ВЧ60) (Appendices 6.A Tables 6.95, 6.95)

The main criteria for the working capacity and the calculation of tooth wave gears are the toughness of the flexible wheel and also the static and dynamic load rating of the wave generator bearings.

6.8.5 Fatigue Strength Calculation of Flexible Wheels

Let us consider an element of the flexible wheel with thickness h and length l (Fig. 6.117) as a part of a circular ring with initial curvature r. After deformation of the flexible wheel by the generator the illustrated element

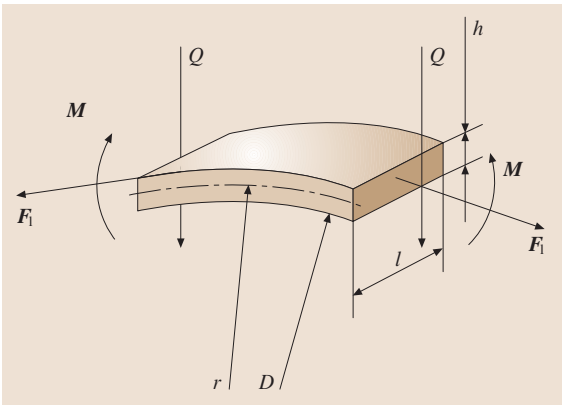


Fig. 6.117 Diagram for the strength calculation of the flexible wheel

has a curvature incrementation that can be represented as depending on the moment of deflection M.

As initial research has shown, the toughness of flexible wheels depends mainly on the stresses that arise as a result of the action of the deflection moment M.

Without taking into consideration the influence of the longitudinal force F_l or transverse force Q the general differential equation of the deflected ring axis has the form

$$\frac{1}{r^2}\left(\frac{\mathrm{d}^2W}{\mathrm{d}\varphi^2} + W\right) = -\frac{M}{EJ}\,, \tag{6.19}$$

where the left-hand side is the ring curvature after deformation, W is the radial displacement, $r = (D+h)/2$ is the radius of the median surface before deformation, and D is the diameter of the flexible wheel opening. On the right-hand side M is the moment of deflection, E is the elasticity module of the wheel material, and J is the moment of inertia of the cross section

$$J = \frac{\ell h^3}{12} = \frac{\ell h^2}{6}\frac{h}{2} = W_b\frac{h}{2}\,,$$

where W_b is the section modulus to bending. The minus sign corresponds to the configuration shown in Fig. 6.117, where moments of deflection that reduce the ring curvature are positive.

Let us transform the right-hand side of (6.19)

$$\frac{M}{EJ} = \frac{M}{EW_b\,(h/2)} = \frac{\sigma_\mathrm{F}}{E\,(h/2)}\,.$$

Calculating the changed equation (6.19) relative to the bending stresses σ_F we obtain

$$\sigma_\mathrm{F} = -\frac{Eh}{2r^2}\left(\frac{\mathrm{d}^2W}{\mathrm{d}\varphi^2} + W\right)\,.$$

Obviously, the bending stresses depend on the deformation law of the flexible wheel. By deformation according to the law $W = W_0 \cos(2\varphi)$, which is similar to that for an ellipse, we have

$$\sigma_F = -\frac{Eh}{2r^2}\left(-4W_0 \cos(2\varphi) + W_0 \cos(2\varphi)\right) .$$

It follows that the bending stresses vary, and that they reach maximum values for $\varphi = 0$ and 90°.

For $\varphi = 0°$

$$\sigma_F = \frac{3}{2}\frac{EhW_0}{r^2} .$$

For $\varphi = 90°$

$$\sigma_F = -\frac{3}{2}\frac{EhW_0}{r^2} .$$

In the general case we can write

$$\sigma_F = A_\sigma \frac{EhW_0}{r^2} ,$$

where A_σ is a coefficient that depends on the form of the deformation. This is particularly so for deformation of flexible wheels with a cam generator with a flexible bearing $A_\sigma = 1.75$.

The characteristics of the cycle of alternating symmetric changes of bending stresses are its amplitude $\sigma_a = \sigma_F$ and mean value $\sigma_m = 0$.

The availability of the gear ring and the tension under the action of the forces F_1 distinguish the real flexible wheel from the smooth ring. Both result in an increase of the acting stresses. Thus, the coefficient K_σ, which takes into account the influence of the gear ring and its tension on the strength of the flexible wheel, is applied to the rated relation ($K_\sigma = 1.5–2.2$; higher values correspond to lower module values and lower values of the rounded radii in the sockets between the teeth).

Upon installation the generator distorts the flexible wheel from only one side. Under the action of the torque the initial form and size of the deformation change in a real gear. This is due to the adjustment of the radial clearance in the supple bearing, the clearances between the bearing cup and flexible wheel, and the contact deformations in the supple bearing and deformations of the stiff wheel. This change in the initial form and deformation size results in an increase of acting stresses, which is taken into consideration through the insertion of the coefficient $K_s = 1.3–1.7$ into the design formula.

The loading of the flexible wheel with torque T and intersecting forces Q that cause the action of the shearing stresses is taken into account by means of insertion of the coefficient $K_\tau = 1.2–1.3$ into the design relation.

Thus, the formula for calculation of the equivalent stresses in flexible wheels has the form

$$\sigma_a = A_\sigma \frac{EhW_0}{r^2} K_\sigma K_s K_\tau .$$

The safety factor according to the fatigue strength of the flexible gear ring is determined from the formula

$$S_F = \sigma_{-1}/\sigma_a , \qquad (6.20)$$

where σ_{-1} is the endurance limit of the material used for the flexible wheel.

The strength condition of the flexible wheel (checking calculation) is

$$S_F \geq [S]_F , \qquad (6.21)$$

where $[S]_F = 1.6–1.7$. Higher values indicate a probability of nonfracture of greater than 99%.

In the case of the design calculation the diameter d of the flexible wheel opening is determined according to the fatigue strength criterion of the flexible ring (Sect. 6.8.6).

Bearing Calculation of Wave Generators

An operational peculiarity of wave generators is the fact that they rotate with high frequency of the input element reacting to high loads of the output elements. The cam wave generator is optimum in terms of load-carrying capacity. The required dynamic load rating of flexible bearings is determined according to the standard method for rolling bearings (Sect. 6.11.14). Wear of the teeth is insignificant and does not limit the gear lifetime in the case of correctly chosen mesh geometry, materials, heat treatment, and lubrication parameters.

6.8.6 Design of Wave Gears

Choice of Mesh Parameters

Tooth Profile. Involute teeth are used in the wave gears, with well-known technological advantages such as the availability of existing tools and the ability to provide sufficiently high multipair toothing under load. To cut involute teeth a tool with a 20° angle of the basic rack profile is used.

It should be noted that the stresses in the rim of the flexible gear wheel reduce when the socket width is increased to a size that is similar to or greater than the tooth thickness. Involute teeth with a wide socket can be cut with a tool with a reduced pitch line depth. The profile of involute teeth with a wide socket is accepted as the basis for the standard series of harmonic reducers for machine-building applications.

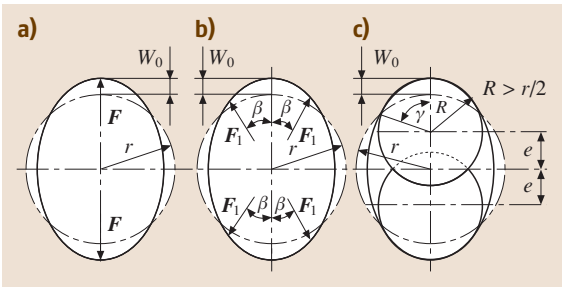


Fig. 6.118a–c Straining diagrams of the flexible wheel with a wave generator. **(a)** With two rollers, **(b)** with four rollers, and **(c)** disk

The construction of the generator determines the deformation form of the flexible wheel as: two-roller (Fig. 6.118a), four-roller (Fig. 6.118b), or disk-shaped (Fig. 6.118c). Any of these forms can be achieved with a *cam generator*. The cam generator retains the set deformation shape better than the others and is thus preferable.

The dimension W_0 of the initial deformation of the flexible wheel is a primary parameter for the calculation of the toothing and generator geometry.

Geometry of the Gear Rings of Flexible Wheels and Rigid Wheels. One of the main geometrical parameters of wave gears is the inner diameter d (mm) of the flexible wheel, the approximate value of which is determined according to *the fatigue strength criterion of the flexible ring*

$$d = 105\sqrt[3]{T/\left(0.16\sigma_{-1}\sqrt{u}/\left(K_\sigma\,[S]_\mathrm{F}\right)\right)}\,,$$

where T is a torque on the low-speed shaft (N m), σ_{-1} is the endurance limit of the flexible steel wheel material (N/mm^2), $K_\sigma = 1.5 + 0.0015u$ is the effective stress concentration factor, u is the transmission ratio, and $[S]_\mathrm{F} = 1.6–1.7$ is a safety factor (higher values of which denote a probability of nonfracture of greater than 99%).

For gears with a cam generator the computed diameter is adjusted according to the outer diameter D of the flexible bearing (Table 6.42).

The width $b_\mathrm{w} = (0.15–0.2)d$ of the gear ring and the thickness S_1 of the flexible wheel are calculated as

$$S_1 = 10^{-4}\left(65 + 2.5\sqrt[3]{u^2}\right)d\,.$$

The circle diameter of the sockets is determined as $d_{fg} = d + 2S_1$. Taking into account that the diam-

Table 6.42 Parameters of flexible ball radial bearings. C_r and $C_{0\mathrm{r}}$ are the dynamic and static load ratings of the bearing, respectively. The ball number $z = 21–23r$ is the bevel dimension

Bearing designation	Dimensions					C_r (kN)	$C_{0\mathrm{r}}$ (kN)	Limit rotation frequency (min^{-1})
	D (mm)	d (mm)	B (mm)	r (mm)	D_w (mm)			
806	$42_{-0.011}$	$30_{-0.010}$	7	0.5	3.969	5.13	5.33	6000
808	$52_{-0.013}$	$40_{-0.012}$	8	0.5	3.969	6.74	7.64	
809	$62_{-0.013}$	$45_{-0.012}$	9	0.5	5.953	10.65	11.98	4980
811	$72_{-0.013}$	$55_{-0.015}$	11	0.5	7.144	13.87	16.83	
812	$80_{-0.013}$	$60_{-0.015}$	13	0.5	7.144	15.48	19.25	
815	$100_{-0.015}$	$75_{-0.015}$	15	1.0	9.128	22.58	28.69	4500
818	$120_{-0.015}$	$90_{-0.020}$	18	1.0	11.113	34.30	46.58	
822	$150_{-0.018}$	$110_{-0.020}$	24	1.0	14.288	51.50	69.02	3480
824	$160_{-0.025}$	$120_{-0.020}$	24	1.0	14.288	53.92	77.00	3000
830	$200_{-0.030}$	$150_{-0.025}$	30	1.0	19.050	92.12	134.38	2520
836	$240_{-0.030}$	$180_{-0.025}$	35	1.5	22.225	121.58	182.91	
844	$300_{-0.035}$	$220_{-0.030}$	45	2.5	28.575	182.33	302.36	1980
848	$320_{-0.040}$	$240_{-0.030}$	48	2.5	28.575	179.10	307.99	
860	$400_{-0.040}$	$300_{-0.035}$	60	2.5	36.513	252.43	502.88	
862	$420_{-0.045}$	$310_{-0.035}$	60	2.5	36.513	252.43	502.88	
872	$480_{-0.045}$	$360_{-0.040}$	72	3.5	44.450	338.45	731.64	

eter d_{fg} is similar to the pitch diameter of the flexible wheel, $d_g \approx d_{fg}$, the module is calculated as $m = d_g/z_g$. It is assumed that $z_g = 2u$ for gears according to (Fig. 6.115a) and $z_g = 2u - 2$ for gear according to (Fig. 6.115b,c).

The value of the module m (mm) is adjusted to the standard one:

Table 6.43 The value of module m (mm) for the wave gearings

First series	Second series
0.25	0.28
0.30	0.35
0.40	0.45
0.50	0.55
0.60	0.70
0.80	0.90
1.00	–

Further, the tooth numbers z_g and z_b are specified, and displacements of the basic rack profile are selected, which provide the diameter d_{fg}. The pitch diameters of the wheels are then calculated: for the flexible one $d_g = mz_g$ and for the stiff one $d_b = mz_b$. The outer diameter of the flexible wheel is determined as $d_{ag} = d_{fg} + 2h_g$, where h_g is the tooth depth of the flexible wheel. When cutting the teeth with a narrow socket on the flexible wheel one has $h_g \approx (1.5–2.0)m$, and when using a wide socket one has $-h_g \approx (1.35–1.55)m$. Then other dimensions of the flexible wheel are set (see below) and the checking calculation is carried out according to the chosen deformation form determining the reserve of the fatigue strength from (6.20) and (6.21).

Design of Flexible Wheels and Rigid Wheels

The Structure of Flexible Wheels. Figure 6.119 shows the most commonly used structures of the flexible wheels of wave gears: (a) with a flexible bottom and a flange for connecting to the shaft, and (b and c) with a spline connection to the shaft. The splines can be cut on the outer (Fig. 6.119b) or inner (Fig. 6.119c) surface of the cylinder. The spline connection reduces the rigidity of the cylinder and, as a result of the axial mobility, decreases stresses in it.

Without the flexible bottom and due to the rigid connection of the cylinder to the shaft (Fig. 6.119d), stresses in the cylinder increase considerably, and the bending rigidity and the resulting load on the generator increase. Such a structure should not be used.

When a flexible wheel is produced as shown in to Fig. 6.119a, the axial yield is provided with a thin bottom at the transfer point of the cylinder to the shaft. *Welded variants* are applied to join the cylinder to the flexible bottom by using a flat butt weld (Fig. 6.119e) or an end joint (Fig. 6.119f). A welded connection of the flexible bottom to the shaft with a size of no more than d_1 (Fig. 6.119g) is also possible; on the flexible bottom an end joint is made according to the shaft diameter.

If the bottom has a flange, the joint can be bolted (Fig. 6.119a), a pin-hole, spline, a keyed, or pressure coupling. Usually a spline connection is used, which allows a relatively small dimension S_6 (Fig. 6.119h).

The shoulder a_1 (Fig. 6.119a) is introduced to reduce the stress concentration at the edges of the gear ring. Towards this end hollow chamfers of major radii

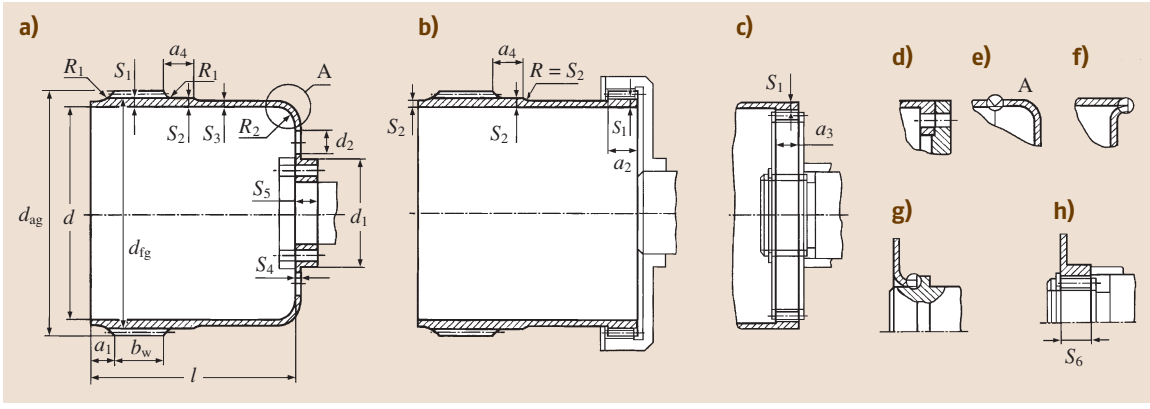


Fig. 6.119a–h Structure of the flexible wheels. (a,d–h) With a flexible bottom, (b,c) with a spline connection to the shaft

R_1 are produced between the gear ring and the cylinder. The openings d_2 increase the compliance of the flexible bottom and provide lubrication circulation. The number and dimensions of the openings are chosen to be as large as possible while still providing sufficient toughness and stability of the bottom.

Previously, d, d_{fg}, d_{ag}, b_w, and S_1 were calculated. Other dimensions given in Fig. 6.119 can be taken according to the following recommendations, which have been tested in practice

$$
\begin{aligned}
d_1 &= (0.6–0.8)d\,, & a_1 &\approx 2S_1\,,\\
l &= (0.8–1.0)d\,, & a_2 &\approx 0.5b_w\,,\\
d_2 &= 0.4(d-d_1)\,, & a_3 &\approx 0.7b_w\,,\\
S_2 &= (0.85–0.9)S_1\,, & a_4 &\approx 0.5b_w\,,\\
S_3 &= (0.6–0.7)S_1\,, & R_1 &\approx 3S_1\,,\\
S_4 &= 1.2S_3\,, & R_2 &\approx 3S_1\,.\\
S_5 &= 3S_1\,,
\end{aligned}
$$

The design of the flexible wheel in Fig. 6.119b is more universal due to the possibility of connection to the shaft or the case.

It is advisable to use a design with a flexible bottom Fig. 6.119a in large-lot production, when metal blocks can be obtained through die forming or expansion. If a method using plastic strain is difficult, welded structures are applied (Fig. 6.119e,f). In single-part production the billet for the flexible wheel as shown in Fig. 6.119a can be obtained by means of turning. However, it is necessary to take into account that the toughness decreases at the same time.

Flexible wheels of tight gears are made in the form of a closed cylinder (Fig. 6.115c), which increases its rigidity considerably, thus increasing the stress level in the cylinder and the load on the generator. To reduce these effects the cylinder length is increased. Transfer

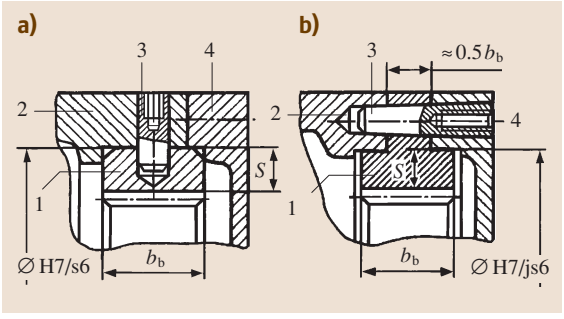


Fig. 6.120a,b Connection with the case of the rigid wheel. **(a)** Without a flange and **(b)** with a flange

of the cylinder to the wall is carried out conically and completed with a thin aperture. The conical area is necessary for ease of installation of the generator with the wheel. The diameter of the flexible wheel d_g and toothing parameters are calculated in the same way as for standard wave gears. The face width is assumed to be $b_w = (0.12–0.18)d$. Other dimensions are assumed to be (Fig. 6.115c)

$$
\begin{aligned}
2l &= (2–1.6)d\,;\\
D_1 &= (1.28–1.35)d\,;\\
\theta &= 1°30'\,;\\
S_3 &= (0.005–0.007)d\,;\\
S_5 &= 1.4S_3\,;\\
S_4 &= 1.6S_3\,.
\end{aligned}
$$

Structure of Stiff Wheels. The stiff wheels of wave gears are similar to the wheels with internal teeth used in standard (with stationary axes) and planetary gears (Figs. 6.95, 6.96, 6.100, and 6.103).

The stiff wheel (1) (Fig. 6.120a) is pressed into the case (2), while the interference fit and three or four pins (3) apply a torque. In the structure shown in Fig. 6.120b the stiff wheel (1) has a flange and centering spigots for mounting the wheel into the case (2) and the cover (4) on the wheel. The wheel structure in Fig. 6.120a is easier to construct, but mounting and dismantling of the stiff wheel are not as convenient. The structure in Fig. 6.120b provides greater wheel rigidity.

The face width b_b of the stiff wheel is designed to be 2–4 mm thicker than that of the flexible one. This allows lower requirements on the positional accuracy of the wheels in the axial direction. The thickness of the stiff wheel is assumed to be $S \approx 0,085d_b$, with the following check on the operating conditions: the maximum radial displacement under load from the toothing forces should not exceed $(0.05–0.02)h$, where h is the tooth setting depth. For involute teeth with a narrow socket one has $h \approx (1.3–1.6)m$, and for teeth with a wide socket one has $h = m$.

Design of Wave Generators

Roller generators are physically simple and easy to manufacture, but they have free areas of the flexible wheel (Fig. 6.118a,b), which do not allow the set deformation form under load to be guaranteed. In view of their small dimensions, bearings for the roller bearers have a limited lifetime, which is why such generators are applied in *lightly loaded* gears.

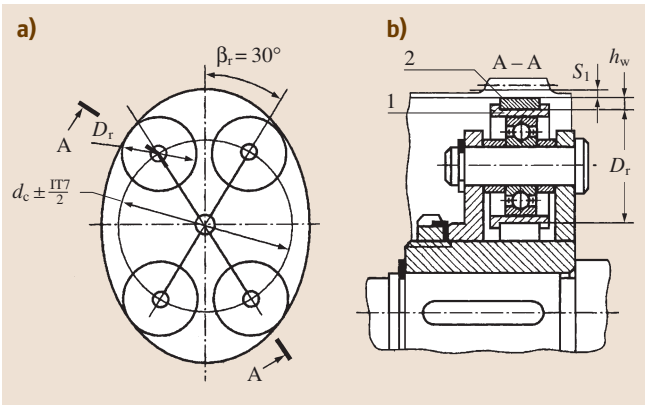


Fig. 6.121a,b Execution version of the four-roller generator

The structure of a four-roller generator is shown in Fig. 6.121. So that the flexible wheel does not flare with the rollers, a jar washer (2) is placed along its inner diameter. In Russia this jar washer is made from the same material as the rollers, e.g., steel 100Cr6 (50–58 HRC) (Appendix 6.A Table 6.95). Furthermore, the jar washer increases the rigidity of the system made up of the flexible wheel and the jar washer and thus reduces distortion of the deformed shape under load. The thickness of the jar washer is assumed to be $h_w \approx 1.5 S_1$. As a roller a rolling bearing, on which the washer (1) is pressed to the ledges, can be used. The ledges are intended to protect the jar washer (2) from axial displacements. The thickness of the washer (1) is assumed to be equal to h_w.

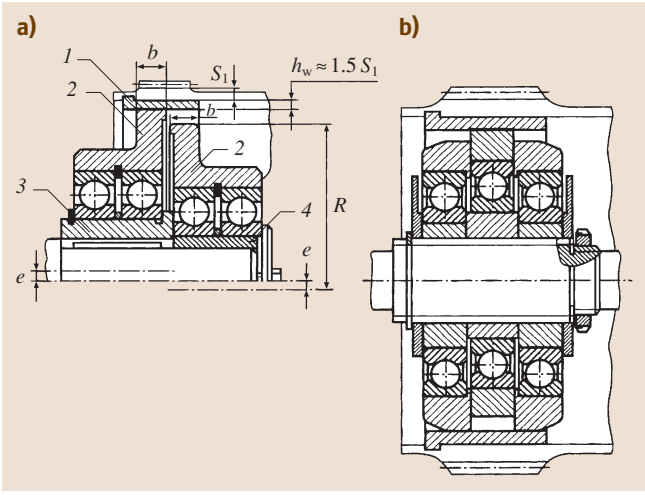


Fig. 6.122a,b Execution version of the generator with disks. **(a)** Two single disks, **(b)** one double disk and one single disk

The diameter of the roller centers is

$$d_c = d + 2W - D_r \,,$$

where d is the inner diameter of the jar washer, W is a radial displacement of the strained flexible wheel at the contact point with the roller, and $D_r \leq 0.33d$.

Disk Generators. The layout of a disk generator is shown in Fig. 6.118c, The structure variants are given in Fig. 6.122. A flexible wheel strained with a generator is positioned along the disk circles on the bow 2γ (Fig. 6.118c), which contributes to keeping the deformation form in the loaded gear. The radii R of the disks and eccentricity e are selected so that the angle γ is 20−40° for the set size of deformation W_0. Usually, $e/W_0 = 3–3.6$, where a lower values means higher γ and lower u.

Each generator disk (2) (Fig. 6.122a) is positioned on two bearings, which protects the disks from warping. The bearings are located on the cylindrical eccentrically positioned shaft journals. The eccentric journals (3 and 4) are formed directly on the shaft and placed on the shaft in the form of bushings. Both bushings are machined as one detail with eccentricity e and a key slot, which is then cut up and fitted onto the shaft. One of its sections is turned through 180°, then the position accuracy of the eccentrics is defined only by the position accuracy of the keys on the shaft. The position accuracy of the eccentrics can be increased by the application of a spline connection with an even number of splines.

As for the roller generator a jar washer (1) is installed to prevent the flexible wheel from flaring. Fixing of the jar washer is difficult in the disk generator because of the axial displacement. In the structure shown in Fig. 6.122a the ledge that fits into the groove of the flexible wheel holds the washer. The ledge height is limited by the allowable value of the elastic tensile strain of the flexible wheel at installation of the jar washer (i. e., it must not exceed it by more than a tenth of a millimeter), which does not guarantee safe blocking of the washer. Moreover, as a stress concentrator, the groove reduces the toughness of the flexible wheel. In Russia the material of the jar washer is steel 100Cr6 (50–58 HRC). The material of the disks is structural steel 45, 40X with quenching of the work surface up to 48–50 HRC (Appendix 6.A Table 6.95).

The offset position along the shaft axis sets up unequal deformation conditions of the flexible wheel in two areas and unbalances the generator load. To decrease this effect the thickness of the disks is reduced

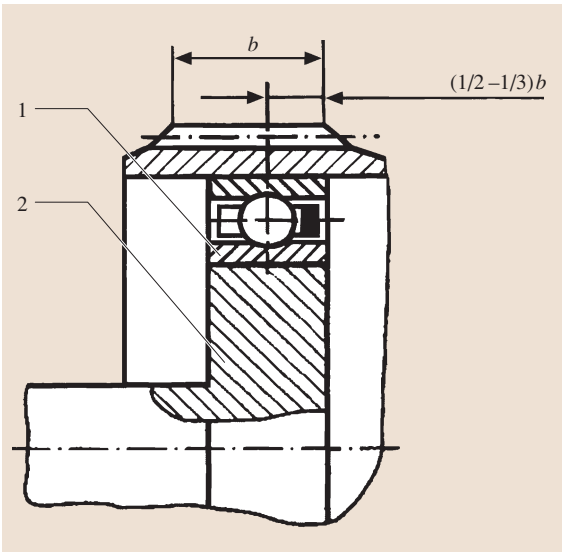


Fig. 6.123 Execution version of the cam generator

to $b \approx 0.1R$. Imbalance in the axial plane can be substantially decreased by applying the generator structure shown in Fig. 6.122b, where one double disk is located symmetrically relative to another single disk. The disk wrap is eliminated through their mutual fit along the ends, which allows installation of each disk on only one bearing. In the structure shown in Fig. 6.122b high execution accuracy of the axial dimensions of the corresponding details is required, the tolerance ranges of which are assigned on the basis of dimension series taking into account axial clearances in the bearings. In power trains the bearings of the disk generator operate under high loads. Thus, the bearing diameter should be chosen to be as large as possible to be in the range of the disk diameter.

Cam Generators. Cam generators consist of an oval cam (2) with a special flexible rolling bearing (1) (Fig. 6.123) pressed on it. The cam profile is designed equidistantly to the accepted deformation form of the flexible wheel, where the initial cam radius is $r = 0.5d$ (Fig. 6.118), where d is the inside diameter of the bearing (Fig. 6.124). The flexible bearing has thin-walled racers, which allows radial deformation of the racers commensurate with their thickness and provides transmission of the rotary motion through the strained racers.

The cam generator retains the deformation shape of the flexible wheel under load better than other generators. With a view to equalizing the load along the tooth

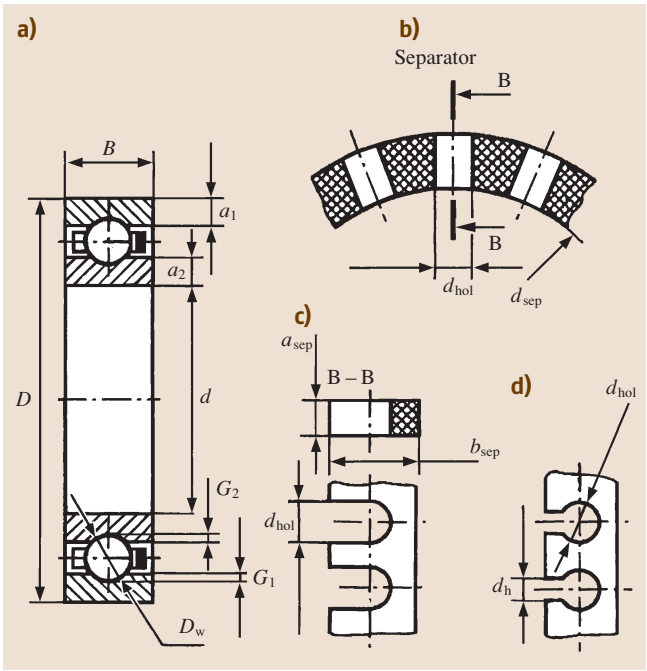


Fig. 6.124 (a) Flexible bearing and (b–d) embodiments of the separator

length and to reduce the axial force on the flexible bearing, the generator is installed in the middle of the gear ring and nearer to the back end.

Flexible bearings (Fig. 6.124a) are distinguished from standard bearings not only by their smaller

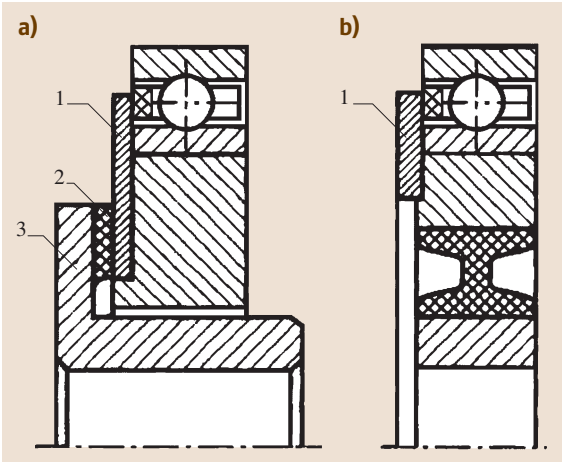


Fig. 6.125a,b Connection of the generator to the shaft by means of an elastic washer (a) and by means an elastic element with increased flexibility under angular warps (b)

racer thickness, but also by the structure of their retainer. The retainer is manufactured from material with a relatively low coefficient of elasticity (textolite, phenilon) with a U-shaped housing. (Fig. 6.124b,c). As a result of the bridge deflections and action of the axial component of the compressive force, under load the retainer is squeezed out of the bearing. It is held, e.g., with a disk (1), fastened to the end of the cam generator (Fig. 6.125). The friction between the retainer and the thrust ring increases losses, which are smaller with a self-closing retainer structure (Fig. 6.124d).

The basic parameters of flexible bearings are (Fig. 6.124):

Table 6.44 Parameters of flexible bearings

Racer thickness	$a_1 \approx a_2 \approx (0.020-0.023)D$
Depth of the racer grooves	$G_1 \approx G_2 \approx (0.05-0.06)D_w$
Inner diameter of the retainer	$d_{sep} = d + 2a_2 + 0.02D + 0.05D_w$
Retainer thickness	$a_{sep} = (0.055-0.060)D$
Retainer width	$b_{sep} = (1.2-1.3)D_w$
Width of the retainer groove	$d_{hol} = (1.01-1.03)D_w$
Housing width	$d_h = D_w$

The parameters of flexible ball journal bearings are given in Table 6.42.

A flexible bearing with opening diameter d is installed on the cam, the diametral dimensions of which are designed to be in the tolerance range $js6$ ($js7$). The outer race of the flexible bearing is mated according to the distance D (Table 6.42) from the inner diameter of the flexible wheel with tolerance range $H7$.

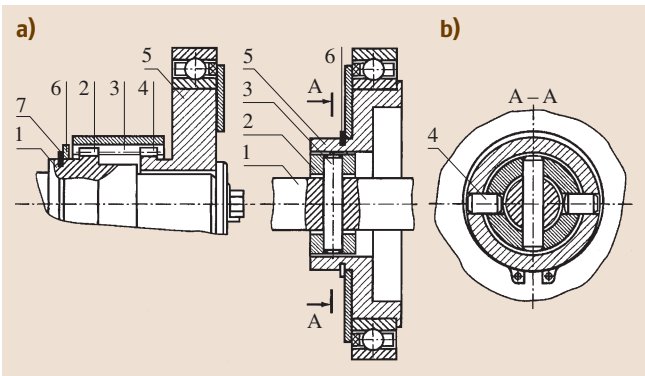


Fig. 6.126a,b Connection of the generator with the shaft by means of the toothed coupling (**a**) of the fixed joint (**b**)

Joining of the Generator to the Shaft. Dead and sliding joints are used for joining the generator to the shaft. In a cam generator with a *dead* joint the cam is installed onto the shaft by means of the usual method, whereby the transmission of the torque is carried through a key joint or spline connection or an interference join. A dead joint can be applied in the case of full coaxiality of the stiff wheel axis and the rotation axes of the generator and the flexible wheel, which can only be achieved with very high requirements of manufacturing accuracy. Deviation from coaxiality results in an uneven load distribution in the toothing zones and an imbalance of power. As a result service life may be reduced and the shaft may even suffer breakage.

For compensation of the kinematic element coaxiality a *sliding* joint can be used to connect the generator to the shaft. This is made with the help of resilient members or fixed joints. In the structure shown in Fig. 6.125a the resilient member is designed in the form of an elastic washer (2), vulcanized to the metal disks (1 and 3), which are further connected to the cam and the shaft. The elastic element shown in Fig. 6.125b possesses increased yielding due to the angular warps. The disadvantage of these joints is the strength reduction of rubber after a certain time.

In the structure shown in Fig. 6.126a the fixed joint is similar to the toothed coupling: the shaft (1) and cam (5) of the generator have gear rings (2 and 4) with external teeth. The washer (6) and the spring ring (7) limit the movement of the bushing (3) in the axial direction.

In the reduction gears swing joint of the generator to the shaft with a cross-shaped position of the pins is used (Fig. 6.126b). The pin (3) goes through the shaft (1) and the bushing (2); two pins (4) go through the bushing (2) and the cam (5). The pins are mounted in the openings with clearances. The internal cam face (5) prevents the pin (3) from falling out, and the spring ring (6) and external shaft face hold the pins (4).

All the structures shown in Figs. 6.125 and 6.126 allow radial and angular cam movements.

For sliding joint the cam generator is the most convenient configuration. For a disk generator, joining of the generator to the shaft with a sliding joint is difficult. In such structures self-adjustment of the elements should be carried out by means of the sliding joint of the rigid wheel with the case or the shaft, which is more difficult and more expensive.

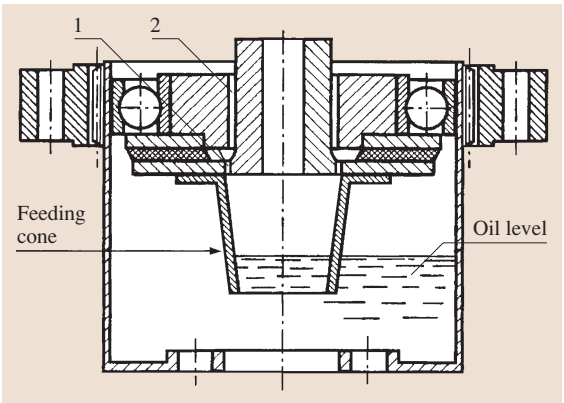


Fig. 6.127 Lubrication through a cone nozzle

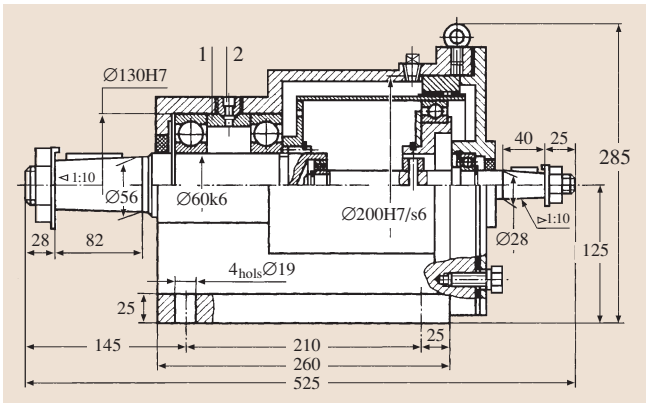


Fig. 6.128 Harmonic reducer

6.8.7 Thermal Conditions and Lubrication of Wave Gears

Thermal conditions of wave gears are calculated in accordance with the known relations (see, e.g., the thermal design of worm-and-wheel gearboxes, Sect. 6.5.13). The allowable oil temperature for reduction gears in industrial applications is $[t] = 70–80°C$. A heat transfer coefficient of as $K = 8–12$ is assumed for small, closed rooms lacking in ventilation. For rooms with intensive ventilation one uses $K = 14–18$, and in the case of fanning $K = 21–30\,\mathrm{W/(m^2\,°C)}$. With a fan installed on the high-speed shaft of the reduction gear, the lower values of this range are used for a rotational frequency of $n \le 1000\,\mathrm{min^{-1}}$, whereas the higher values for K apply for rotational frequencies of $n \ge 2800\,\mathrm{min^{-1}}$.

For reduction gears used in machine-building applications fluid mineral oil with kinematic viscosity of $68–100\,\mathrm{mm^2/s}$ is recommended. If needed, a semisolid lubricant is used. The bearings of the generator and the toothing are lubricated by mounting the reduction gear and is occasionally done online. To change the semisolid lubricant takes approximately 1000 working hours.

A semisolid lubricant can be used in the case of vertical installation of the reduction gear axis. A special oil feeding unit is mounted in the reduction gear for liquid oil lubrication (Fig. 6.127). Under the action of centrifugal forces oil rises inside the *feeding cone*, goes through the openings (1) and clearance (2) into the generator and thereby reaches the bearing and toothing. The structure in Fig. 6.127 is recommended for rotational frequencies of $n \ge 960\,\mathrm{min^{-1}}$.

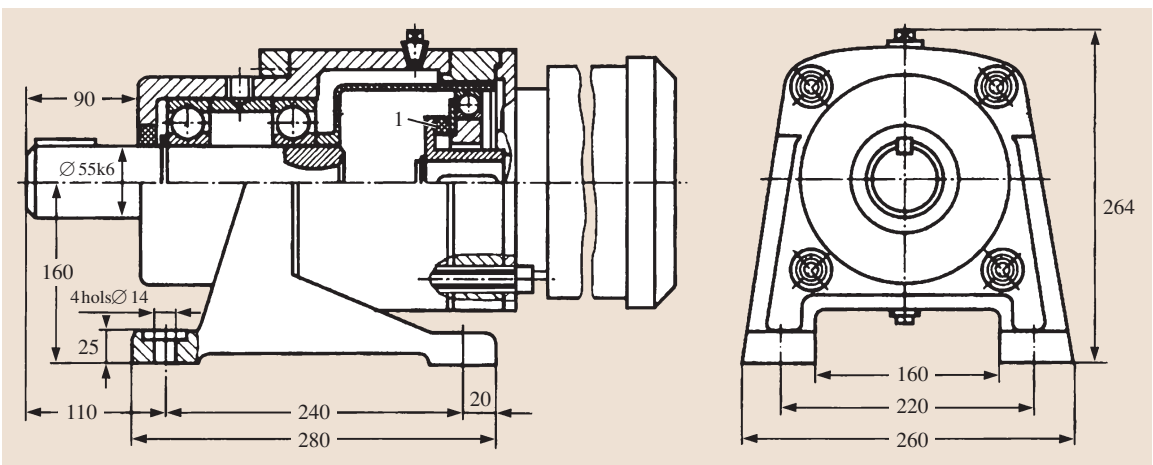


Fig. 6.129 Harmonic motor-reducer

It is recommended to use a quantity of oil in the reduction gear such that its level reaches the center of the lower ball of the flexible bearing when the reduction gear is in a horizontal position. For $n < 960\,\mathrm{min}^{-1}$ and vertical shaft installation it is possible to fill the reduction gear with oil completely.

6.8.8 Structure Examples of Harmonic Reducers

Wave gears are standardized, with the internal diameter of the flexible wheel ranging from 50 to 250 mm, torques from 35 to 5600 N m, and transmission ratios from 76 to 275.

Figure 6.128 shows a typical structure of a harmonic reducer. Distinctive features of the structure are: the double-seat shaft of the generator, the junction of the cam generator with the shaft with an articulated shaft coupling (Fig. 6.126b), the welded connection of the flexible wheel cylinder to the bottom, the spline connection of the flexible wheel to the shaft, the axial fixation of the output shaft bearings in the case with spacing bushing (1) and three adjusting screws (2), the joint with interference of the stiff wheel with the case, and the cylindrical form of the internal housing without inner valleys and pockets, which simplifies casting and cleaning after casting and machining.

Figure 6.129 shows a harmonic reducer with removable legs mounted to the cylindrical case with screws. Distinctive features of this reducer are: the cantilever position of the generator on the electric motor shaft, the junction of the generator with the shaft with the vulcanized elastic washer (1); the flexible wheel is pressed with the following machining, the joint with interference of the flexible wheel to the shaft, and the fixation of the stiff wheel on the case with screws and pins.

6.9 Shafts and Axles

6.9.1 Introduction

Gear wheels, pulleys, chain wheels, and other revolving machine parts are installed onto shafts and axles. A *shaft* is intended for transmission of the torque along its axis, for holding the details that are located on it, and for withstanding the action of the forces acting on these details. An example is the shaft of a reduction gear (Fig. 6.130). Under operation the shaft experiences the action of *bending and torsion* stresses, and in some cases extra tension or compression stress [6.70–77].

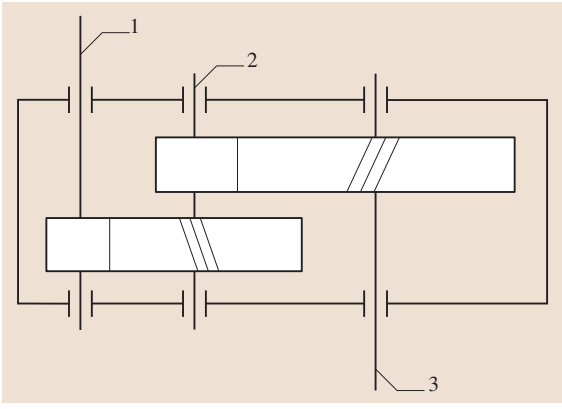


Fig. 6.130 Diagram of the cylindrical double-reduction gear unit

An *axle* only supports the details installed on it and takes the forces acting on these details. An example is the axle of a railroad car (Fig. 6.131). In contrast to a shaft an axle does not transmit torque and therefore *does not experience torsion*. An axle can be *fixed* or *can rotate* with the details placed on them. Rotating axles provide a better operating environment for the bearings. Fixed axles are cheaper but require integration of the bearings into the details rotating on the axles.

Most shafts have an invariable nominal axle geometry, i. e., they are *stiff* shafts. A special group includes *flexible* shafts that have a variable form of the geometrical axles. Depending on the form of the geometrical axles, shafts are divided into *straight* (Fig. 6.132) and *indirect* crankshafts, which serve to transform alternating motion into rotational motion (or vice versa), and *eccentric* crankshafts.

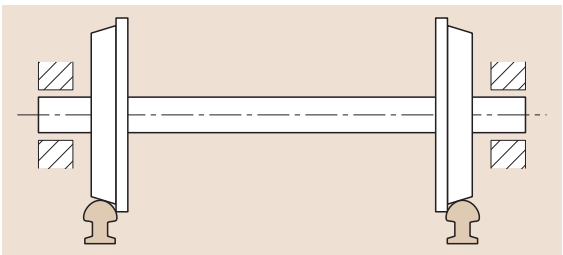


Fig. 6.131 Axles of a railroad car

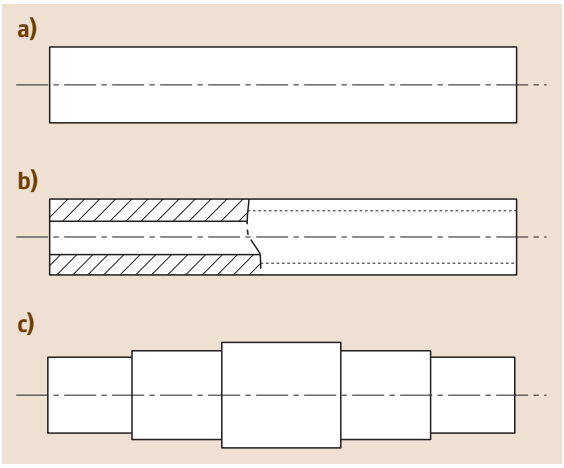


Fig. 6.132a–c Straight shaft. **(a,b)** Single-diameter part, **(c)** multidiameter shaft

As a rule, the axles used are straight. Straight shafts and axles have the form of bodies of revolution and differ little from each other in design. Straight shafts and axles can be of constant diameter: *single diameter* (Fig. 6.132a,b) or *multidiameter* (most shafts, Fig. 6.132c).

Depending on the form of the cross section there are *solid* and *hollow* shafts, and axles (with axial opening, Fig. 6.132b). Hollow shafts are used for mass reduction, as well as when other details and materials must be passed through the shafts or positioned inside the details or materials (oils, cooling gases, and liquids).

Depending on the outer outline of the cross section the shafts are divided into a *spline* and a *key*, which have a spline profile or a profile with a key groove along a certain length. Shafts are also classified according to conventional features, e.g., according to the relative speed of rotation in the unit (e.g., in the reduction gear in Fig. 6.130) as high-speed 1, medium-speed 2, and low-speed 3, or according to position in the unit: input 1 (drive), countershaft 2, and output 3 (driven). *Journals* serve as supports for shafts and axles. Intermediate journals are called *necks*.

Shaft Form According to Length

Based on data of uniform strength it is advisable to design shafts similar to bodies with equal bending resistance in their longitudinal section, outlined by a cubic parabola. *Multidiameter shafts* approximate to the form of a body with equal resistance. This form also simplifies manufacture and mounting of the details on the shaft.

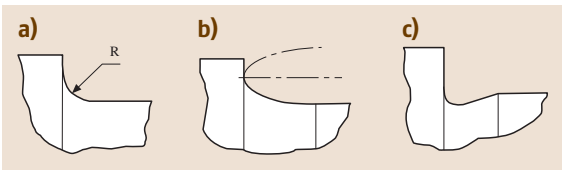


Fig. 6.133a–c Transition shaft sections. **(a)** With a hollow chamfer of the constant and **(b)** varying radius, and **(c)** with a groove with rounding

Transitional sections of the shafts and axles between two stages of different diameters are designed as follows. For a hollow chamfer of constant radius (Fig. 6.133a), a hollow chamfer is a surface of the graded junction from the light section to the major one; for a hollow chamfer of varying radius (Fig. 6.133b), with a groove with a fillet for outlet of the grinding wheel is used (Fig. 6.133c). Transitional sections are stress concentrators. An effective method for stress concentration reduction in the transitional sections is to increase their yielding (e.g., by means of increasing the hollow chamfer radii and introducing balancing grooves). Strain hardening (cold hardening) of the hollow chamfers increases the load-carrying capacity of shafts and axles.

Shaft and Axle Materials

Carbon and alloy steels (Table 6.45) are the basic materials for shafts and axles due to their toughness, their high elasticity coefficient, and the strengthening ability and availability of the required cylindrical blanks through forging or rolling.

In Russia for shafts and axles subjected to a rigidity criterion and not heat-treated, steel grades E 295 and E 355 (EN) are mainly used. Heat-treated medium carbon and alloy steel grades 35 (GOST) [6.78–88], C 46, C 45 (EN), and 37 Cr 4 (DIN) are used for most shafts. Alloy steel grades 40 NiCr 6 (DIN), 40 NiCrMo 4 KD (EN), 30 MnCrTi (DIN), 20 CrMo 5 (DIN), and others are used for high-duty shafts of critical machines (Appendix 6.A Table 6.95). Shafts made from these steels are usually refined, quenched with a high-temperature tempering, or surface-hardened by heating with RF current and low-temperature tempering (spline shafts).

Alloy steels are used following certain constructive reasons, e.g., the need to provide the required contact or bending strength of the teeth cut directly on the shaft, or the need to obtain the required qualitative characteristics of the surface layers on individual shaft areas.

For high-speed shafts rotating in a friction bearing, high journal hardness is required. In this case, shafts are

Table 6.45 Mechanical characteristics of steels used in the production of shafts

Steel grade	Blank diameter (mm)	Hardness (HB)	Mechanical characteristics (N/mm^2)					Coefficient ψ_τ
			σ_t	σ_y	τ_y	σ_{-1}	τ_{-1}	
Ст5	Any	≥ 190	520	280	150	220	130	0.06
35	Any	≥ 207	540	320	160	270	160	0.07
45	Any	≥ 200	560	280	150	250	150	0.05
	≤ 120	≥ 227	820	640	290	360	200	0.09
	≤ 80	≥ 260	940	760	390	410	230	0.10
40X	Any	≥ 200	730	500	280	320	200	0.05
	≤ 200	≥ 240	790	640	380	370	210	0.09
	≤ 120	≥ 270	980	780	450	410	240	0.10
40XH	Any	≥ 240	820	650	390	360	210	0.05
	≤ 200	≥ 270	980	785	450	420	230	0.10
20X	≤ 120	≥ 197	650	400	240	310	170	0.07
12XH3A	≤ 120	≥ 260	930	685	490	430	240	0.10
18XГT	≤ 60	≥ 330	1180	930	660	500	280	0.12
30XГT	Any	≥ 270	950	750	520	450	260	0.05
	≤ 120	≥ 320	1150	950	665	520	310	0.12
	≤ 60	≥ 415	1500	1200	840	650	330	0.15
20X2H4A	≤ 200	≥ 321	1270	1080	740	550	330	0.12

produced from cemented steel grades 20CrS4 (DIN), 14NiCr10 (5732) (DIN), 20MnCr5G (DIN) or nitrided steel grades 40NiCrMo4KD (EN), 41CrAlMo7 (EN) (Appendix 6.A Table 6.95). Chrome-plated shafts have the greatest wear resistance. Based on experience in automobile construction, chromium plating of crankshaft journals increases the lifetime of regrinding by three to five times.

For shafts where the dimensions are defined by the rigidity criterion, it is advisable to apply only strong heat-treated steels in order to guarantee the service life of splins and other wearing faces.

For the production of shaped shafts (crankshafts, shafts with major flanges and openings, and other heavy shafts) high-duty cast irons (with globular graphite), and inoculated cast irons are used along with steel. The lower toughness of iron shafts is substantially compensated for by the more exact forms of the shafts (especially crankshafts), the decreased stress concentration sensibility of iron, lower sensibility (because of lower elasticity coefficient) to inaccurate bearing positions in multiple-bearing shafts, and lower additional dynamic load due to their greater damping capacity. Long hollow shafts are sometimes manufactured from composite materials by means of multilayer winding of a band on a straight plug.

Rounds are used as billets for steel shafts with a diameter of up to 150 mm, and forgings are used for shafts with large diameter and shaped shafts. Shafts are sometimes welded from tubes or sheets with welded or set flanges. Shafts are subjected to turning and mounting surfaces are ground. High-duty shafts are ground over their entire surface.

6.9.2 Means of Load Transfer on Shafts

The main forces acting on shafts are forces from the gears. The forces on the shafts are transmitted through the details set on them: gear or worm wheels, pulleys, chain wheels, half-couplings, etc.

The three-dimensional diagram in Fig. 6.134 shows the forces loaded on the shafts of a double-reduction cylindrical gear unit with helical gearing. In analytical models these forces, as well as the torques, are shown as concentrated, applied at the center of the hubs (Fig. 6.135). The influence of the weight of the shaft and the installed details is neglected (with the exception of heavy flyweights and so on). Frictional forces in the bearings are not taken into account.

Transmission of the torque is carried out through junctions: with interference, spline, key joints, frictional bevel rings, etc. In joints with interference, cylindrical elements are mainly used because they are easier

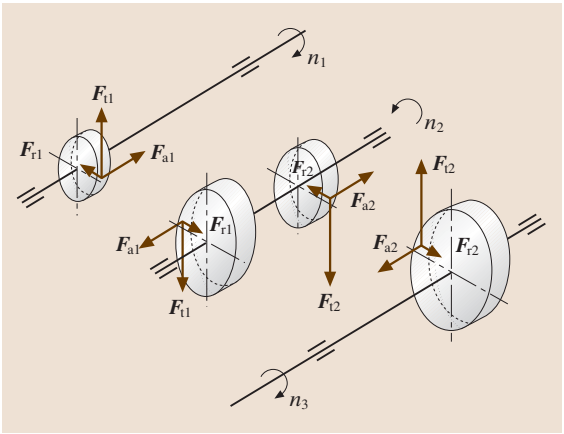


Fig. 6.134 Diagram of the forces loading the shafts of a double-reduction gear unit

to produce. Bevel connections are applied for ease of installation on the shaft and removal of heavy details, for quick detail replacement, e.g., changeable pinions, for the provision of the required interference, and to increase the accuracy of detail centering. Bevel connections are mostly produced on the end sections of the shafts; the required axial force is provided by a nut or a screw and face washer.

Radial forces are transmitted through direct contact of the hub set on the shaft (the most common case), or through bearings (crankpins of the crankshafts). Axial forces are transmitted as follows: those that are substantial in size are transmitted through the detail stop into the shoulders of the shaft (Fig. 6.136a) or through interference fitting of the details; medium forces are

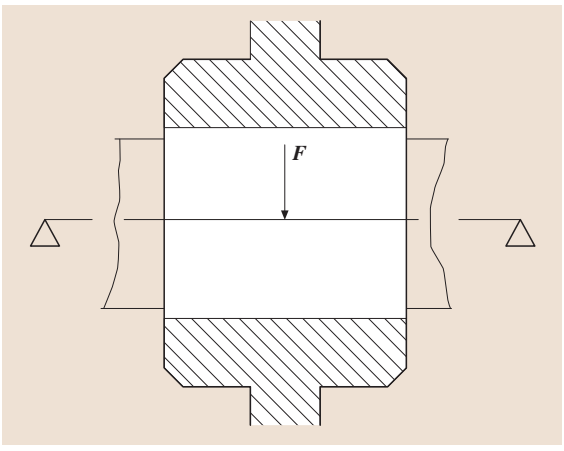


Fig. 6.135 Transmission diagram for the radial force from the wheel hub to the shaft

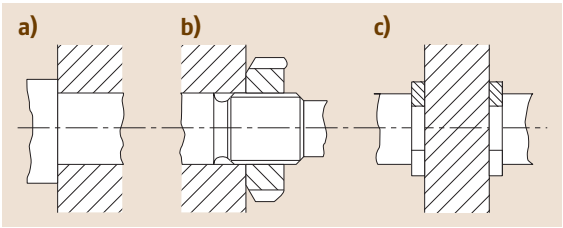


Fig. 6.136a–c Transmission methods of the axial force on the shaft. **(a)** With a stop into the ledge, **(b)** with a nut, **(c)** with spring planar thrust rings

transmitted through nuts (Fig. 6.136b) or spring planar thrust rings (Fig. 6.136c); and light forces are transmitted through spring rings and stop screws.

6.9.3 Efficiency Criteria for Shafts and Axles

The main efficiency criteria for shafts and axles are strength and rigidity. If necessary, the shafts are calculated on chatter stability. For strength analysis of shafts and axles, the bending moment and torque, together with longitudinal force diagrams, are used. In operation, shafts and rotating axles experience the action of cyclically changing stresses. Strength is estimated by using the safety factors S_y in the static strength analysis of shafts and axles, and with S for fatigue strength analysis; rigidity is estimated by the deflection, angular deflections, or torsion of the sections at the installation points of the details. It is known from practice that the fracture of shafts and axles of high-speed machines is in most cases due to fatigue, so fatigue strength analysis is the main consideration.

The most important force factors in shaft and axle design are the torques T and the moment of deflection M. The influence of stretching and compressive forces on strength is rarely taken into account.

6.9.4 Projection Calculation of Shafts

The projection calculation of shafts is carried out based on the static strength in order to determine the approximate diameter of the single grades. At the beginning of the calculation only the torque T is known. It turns out that the moments of deflection M can be calculated only after the shaft design has been determined, when the length and application sites of the acting loads have been determined by the system setup. This is why the projection calculation for shafts is carried out based only on torsion, and the influence of bending, stress con-

Table 6.46 Allowable stresses $[\tau]_t$ for the calculation of the diameter of the end shaft section (**a** under constant load and of constant direction, **b** under varying load, if maximum runs up to the double value, **c** in the case of pure torsion with varying direction. The radial force F is applied to the middle length of the shaft end)

Tensile stress of the shaft material σ_t (N/mm^2)	Hardness HB	Values τ_t (N/mm^2)						
		Pure torsion			Torsion and bending from radial force F			
					$F \le 250\sqrt{T}$		$F > 250\sqrt{T}$	
		a	*b*	*c*	*a*	*b*	*a*	*b*
500–850	145–250	40	28	20	28	20	14	10
> 850–1200	> 250–350	56	40	28	40	28	20	14
> 1200	> 350	$\frac{80}{112}$	56	40	56	40	28	20

centration, and the pattern of stress change on the shaft strength is compensated for by increasing the allowable stress $[\tau]_t$ on the torsion.

In the projection calculation for reduction gear shafts the cross-sectional diameter of the typical area is usually determined at the end of the input (1) and output (3) shaft and the installation position of the gear wheel on the countershaft (2) (Fig. 6.130). The diameters of other areas are assigned by working out the shaft design, taking into account functionality, fabrication technique, and assembly.

The diameter d (mm) of the rated shaft cross section is determined from the formula

$$d \ge 10\sqrt[3]{T/(0.2\,[\tau]_t)}\,,$$

where T is the torque acting in the rated shaft cross section (N m) and $[\tau]_t$ is the allowable stress on the torsion (N/mm^2).

The values of the allowable stresses on the torsion $[\tau]_t$ for the determination of the diameter d of the input (output) shaft end are given, depending on the load condition, strength, and material hardness, according to Table 6.46.

A sketch to work out the shaft design is then made, specifying its form and dimension after the choice and calculation of the bearings, the calculation of the connections that take part in the driving torque transmission, and the design of the structural components resulting from the chosen fixation methods and adjustment of the axial position of the installed on the shaft details, of the shaft itself in the case, as well as the machining technique for the individual areas.

6.9.5 Checking Calculation of Shafts

After a complete design embodiment of the shaft, a checking calculation for the static strength, the fatigue strength, and the rigidity is carried out. In the analytical model, the shafts are regarded as bars on rigid pivoted mounts.

By the choice of the bearing type it is supposed that the shaft deformations are small and, if the bearing allows at least a slight tilt or journal displacement (e.g., in the range of the clearances between the rolling elements and the rings), it is considered a pivoted mount, either *hinged immovable* or *hinged movable*. Rolling or friction bearings simultaneously taking radial and axial forces are considered as hinged immovable (*fixed*) supports (Fig. 6.137a–c), and bearings taking only radial forces are considered hinged movable (*floating*) supports.

The relative support is positioned in the middle of the width of the radial rolling bearings (Fig. 6.137a) or at a displacement a from the face for radial-thrust bearings (Fig. 6.137b). For tapered roller bearings $a = 0.5[T + (d + D)e/3]$, where T is the mounting height, d is the opening diameter, D

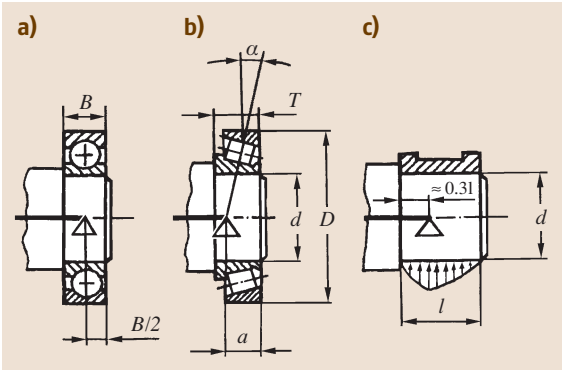


Fig. 6.137a–c Position of the relative support for installation of the bearing: **(a)** radial, **(b)** radial-thrust, and **(c)** friction bearing

is the outer diameter, and e is the axial loading factor.

In shafts rotating with non-self-installing friction bearings, the stress along the length l of the bearing is distributed unevenly as a result of shaft deformation. Thus, the relative pivoted mount is displaced in the direction of the loaded spacing (Fig. 6.137c), on the condition that $0.3l \leq 0.5d$.

Calculation Procedure

The calculation is carried out in the following manner: the analytical model is diagramed according to the drawing of the assembling shaft unit (Fig. 6.138), on which all the external forces that load the shaft are plotted, locating the planes of their action on two mutually perpendicular planes (horizontal X, vertical Y). Then the reaction at the supports is determined in the horizontal and vertical planes. In the same planes the bending moment M_x and the M_y diagram is plotted. The twisting moment diagram M_t, and the longitudinal force F_a diagram are plotted then separately. At the application points of the external bending moments the rated bending moments are determined to the left and to the right of the section.

The cantilever area of the output (input) shaft can be loaded with the radial force F_c (Fig. 6.138) acting from the side of the joint sleeve, belt, or chain drive. If the direction of the force vector F_c is not known beforehand, the bending moment M from this force is plotted separately, without superposing it onto the planes X and Y.

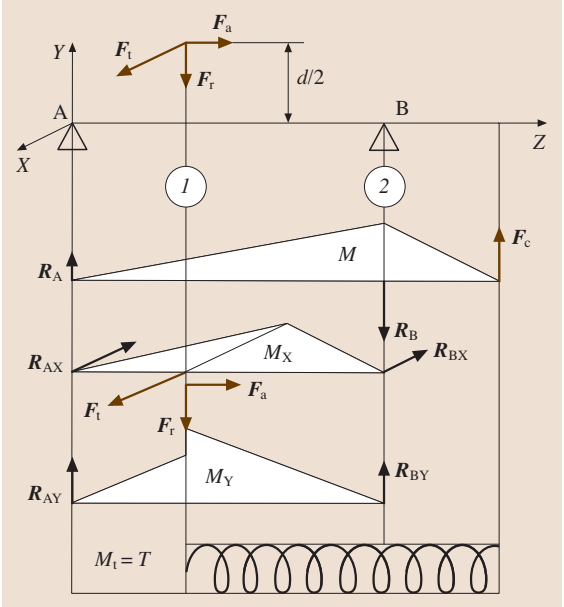


Fig. 6.138 Diagram for the strength calculation of a shaft

Weak sections are ascertained on the basis of the moment diagrams, the dimensions and form of the shaft cross sections, and the availability of the stress concentrators. Two to three sections that are loaded with moments are usually selected, where external forces, moments, support reactions, or form change points along the shaft length are applied.

Weak sections for the analytical model in Fig. 6.138 of the output shaft of the reduction gear are the following:

- Section 1, at the point of the gear wheel installation, which is loaded vt the twisting moment M_t and the bending moments M_x, M_y, and M. Potential stress concentrations are located at the interference fit of the wheel hub onto the shaft, key groove, and splines.
- Section 2, at the point of the rolling bearing installation, which is loaded with the twisting moment M_t and the bending moment M. The stress concentrator is an interference fit of the bearing cone onto the shaft.

Here one should bear in mind that the shaft diameter in section 2 is, as a rule, less than the shaft diameter in section 1. The shaft strength is checked in these weak sections.

Static Strength Calculation

Checking of the static strength is carried out with a view to prevent plastic strains during the action of short-time overloads (e.g., during run-up, acceleration, speed reversal, braking, or response of the security device).

The overload value is calculated by taking into account the specific character of the machine operation according to the starting moment of the electric motor, the limiting moment on release of the overload clutch, the inertial moments caused by sudden braking, etc.

The overload factor $K_0 = T_{max}/T$ is used in the calculation, where T_{max} is a maximum short-term acting torque (overload moment) and T is the nominal (rated) torque.

The normal stresses σ (N/mm^2) and shearing stresses τ (N/mm^2) are determined from the calculation in the shaft section concerned under maximum loads

$$\sigma = 10^3 M_{max}/W + F_{max}/A \,,$$

$$\tau = 10^3 M_{t\,max}/W_t \,,$$

where $M_{max} = K_0(\sqrt{M_x^2 + M_y^2} + M)$ is a total bending moment, N m; $M_{t\,max} = T_{max} = K_0 T$ is a twisting moment (N m), $F_{max} = K_0 F_a$ is an axial force (N), W and W_t are modules of the shaft section in the bending and torsion calculation (mm^3), and A is a cross-sectional area (mm^2).

The partial load factor according to the normal $S_{y\sigma}$ and shearing $S_{y\tau}$ stresses is

$$S_{y\sigma} = \sigma_y/\sigma\ ; \quad S_{y\tau} = \tau_y/\tau\ ,$$

where σ_y and τ_y are the yield strengths of the shaft material under bending and torsion (N/mm^2) (Table 6.45).

The general load factor according to the yield strength in the case of joint action of the normal and shearing stresses is

$$S_y = S_{y\sigma} S_{y\tau} / \sqrt{S_{y\sigma}^2 + S_{y\tau}^2}\ .$$

Static strength is provided if $S_y \geq [S]_y$. The minimum allowable value of the reserve coefficient in the yield strengths is assumed to be in the range $[S]_y = 1.3–2.0$, depending on the importance of the structure and the consequences of shaft fracture, accuracy determination of loads and stresses, the level of the production technique and control, and the similarity and stability of the material.

Fatigue Strength Calculation

Under the action of environmental stresses the rotating shafts are subjected to *periodic loading*. A short-cut calculation for *regular loading* (with constant parameters of the loading cycle during the entire running time) is given below. More precise calculations according to the revised summation theory of the damage by irregular loading are given in specialized literature.

As a consequence of the shaft rotation, bending stresses at different points of its cross section change according to the completely reversed cycle (Fig. 6.139a) with the following parameters: σ_a is a stress cycle amplitude and $\sigma_m = 0$ is a mean stress. Torsion stresses are proportional to the torque and change according to the zero-to-tension stress cycle (Fig. 6.139b) with the following parameters: τ_a a stress cycle amplitude, τ_m a mean stress, and $\tau_a = \tau_m$. The choice of the zero-to-tension stress cycle for the torsion stresses is based on the fact that the shafts transmit running torques that vary in value but which are constant in direction.

In the short-cut fatigue strength calculation the parameters of the cycle are rated according to the maximum of the long-term load. The calculation is carried out in the checking form of the assurance factor S in the sections presumed to be weak, as previously determined in accordance with the shaft form, bending moment diagram, and the position of the stress concentration zones.

Strength is guaranteed if $S \geq [S]$. The minimum allowable value of the load factor is $[S] = 1.5–2.5$. A value within this range is assumed, depending on the degree of certainty of the effective load determination and taking into account the importance of the structure, on the empirical basis of the previous calculations and control over the behavior of the machine in operation.

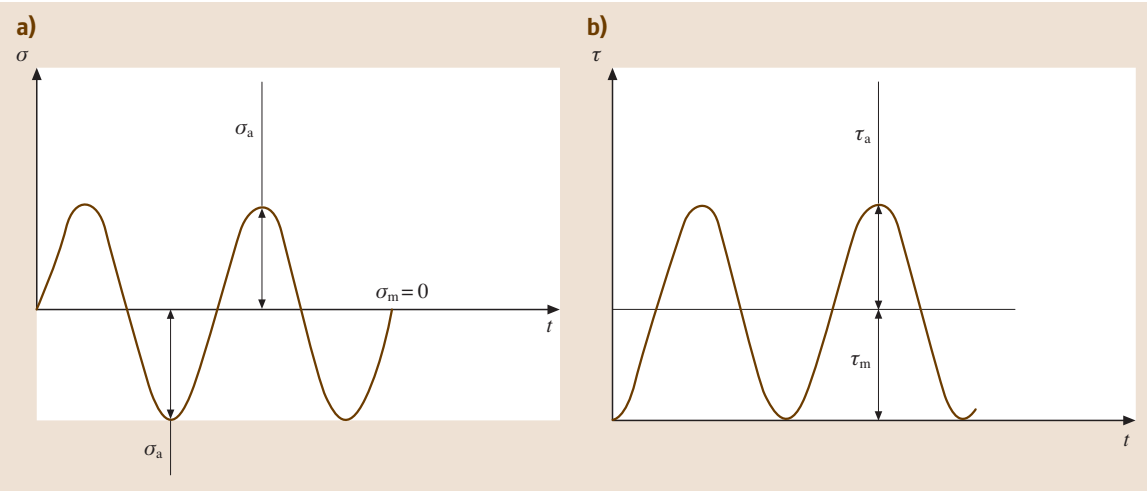


Fig. 6.139a,b Planning of a change of (a) bending stresses and (b) torsion stresses

Table 6.47 Influence factors $K_{d\sigma}$ and $K_{d\tau}$ of the absolute dimensions of the shaft cross-section

Stress condition and material	$K_{d\sigma}(K_{d\tau})$ with shaft diameter d (mm)					
	20	30	40	50	70	100
Bend for carbon steel	0.92	0.88	0.85	0.81	0.76	0.71
Twist for all the steels and bend for alloy steel	0.83	0.77	0.73	0.70	0.65	0.59

Table 6.48 Influence factors $K_{F\sigma}$ and $K_{F\tau}$ of the finished treatment

Type of machining	Roughness parameter Ra (μm)	$K_{F\sigma}$ for σ_t (N/mm^2)		$K_{F\tau}$ for σ_t (N/mm^2)	
		≤ 700	> 700	≤ 700	> 700
Fine grinding	≤ 0, 2	1	1	1	1
Fine turning	0.2–0.8	0.99–0.93	0.99–0.91	0.99–0.96	0.99–0.95
Finish grinding	0.8–1.6	0.93–0.89	0.91–0.86	0.96–0.94	0.95–0.92
Finish turning	1.6–3.2	0.89–0.86	0.86–0.82	0.94–0.92	0.92–0.89

For each of the fixed presumably weak sections the general load factor S is calculated as

$$S = S_\sigma S_\tau / \sqrt{S_\sigma^2 + S_\tau^2} \geq [S] \ ,$$

where S_σ and S_τ are safety factors for the normal and shearing stresses

$$S_\sigma = \sigma_{-1D}/\sigma_a \ ,$$
$$S_\tau = \tau_{-1D}/(\tau_a + \psi_{\tau D}\tau_m) \ .$$

Here σ_a and τ_a are stress amplitudes, τ_m is a mean stress (Fig. 6.139), and $\psi_{\tau D}$ is an influence factor of the stress cycle unbalance for the shaft section concerned.

Stresses in the weak sections are determined from the formulas

$$\sigma_a = 10^3 M_c / W \ ,$$
$$\tau_a = 10^3 M_t/(2W_t) \ , \quad \tau_m = \tau_a \ ,$$

where $M_c = (\sqrt{M_x^2 + M_y^2} + M)$ is a resultant bending moment (N m), M_t is a twisting moment ($M_t = T$) (N m), and W and W_t are modules of the shaft section for bending and twisting (mm^3).

The influence factor $\psi_{\tau D}$ of the stress cycle imbalance for the shaft section concerned is

$$\psi_{\tau D} = \psi_\tau / K_{\tau D} \ ,$$

where ψ_τ is a sensitivity index of material to the stress cycle unbalance (Table 6.45).

The endurance limits of the shaft in the section are

$$\sigma_{-1D} = \sigma_{-1}/K_{\sigma D} \ ,$$
$$\tau_{-1D} = \tau_{-1}/K_{\tau D} \ ,$$

where σ_{-1} and τ_{-1} are the endurance limits of the smooth specimens in the completely reversed cycle of bending and twist (Table 6.45), and $K_{\sigma D}$ and $K_{\tau D}$ are reduction factors for the endurance limit.

The values $K_{\sigma D}$ and $K_{\tau D}$ are determined from the relations

$$K_{\sigma D} = (K_\sigma / K_{d\sigma} + 1/K_{F\sigma} - 1)/K_v \ ,$$
$$K_{\tau D} = (K_\tau / K_{d\tau} + 1/K_{F\tau} - 1)/K_v \ ,$$

where K_σ and K_τ are effective stress concentration factors of bending and twist. The influence on the endurance limit of the shaft form change in the axial direction or the cross section is also taken into account (transition area, key groove, splines, thread, etc.). Pressure at the installation point of the details mounted with interference (gear wheels, rolling bearings) is also a stress concentrator. Stress concentration decreases the endurance limit. $K_{d\sigma}$ and $K_{d\tau}$ are influence factors of the dimensions of the absolute cross-section (Table 6.47). The higher the absolute dimensions of the cross section of the detail, the lower the endurance limit. $K_{F\sigma}$ and $K_{F\tau}$ are influence factors of the surface finish (Table 6.48). With increasing surface roughness the endurance limit of the detail is lowered. The development of corrosion considerably reduces the endurance limit during operation. K_v is an influence factor of surface hardening (Table 6.49). Surface hardening of the detail increases the endurance limit considerably. Surface hardenings are more effective than volumetric ones, which are often accompanied by impact strength reduction and an increase in the stress concentration sensitivity. For example, case-hardening and quench-hardening increase the fatigue strength by 30–40% or more in comparison with volume quenching for the same hardness.

The values of the coefficients K_σ and K_τ are taken from tables. For a step junction with a hollow chamfer

Table 6.49 Influence factors K_V of surface hardening

Type of surface hardening of the shaft	The values K_V with		
	$K_\sigma = 1.0$	$K_\sigma = 1.1-1.5$	$K_\sigma \geq 1.8$
Quenching with RF current	1.3–1.6	1.6–1.7	2.4–2.8
Nitriding	1.15–1.25	1.3–1.9	2.0–3.0
Roller knurling	1.2–1.4	1.5–1.7	1.8–2.2
Cloud burst hardening	1.1–1.3	1.4–1.5	1.6–2.5
Without hardening	1.0	1.0	1.0

Table 6.50 Effective stress concentration factors K_σ and K_τ for a step junction with a hollow chamfer

t/r	r/d	K_σ with σ_t (N/mm^2)				K_τ with σ_t (N/mm^2)			
		500	700	900	1200	500	700	900	1200
2	0.01	1.55	1.6	1.65	1.7	1.4	1.4	1.45	1.45
	0.02	1.8	1.9	2.0	2.15	1.55	1.6	1.65	1.7
	0.03	1.8	1.95	2.05	2.25	1.55	1.6	1.65	1.7
	0.05	1.75	1.9	2.0	2.2	1.6	1.6	1.65	1.75
3	0.01	1.9	2.0	2.1	2.2	1.55	1.6	1.65	1.75
	0.02	1.95	2.1	2.2	2.4	1.6	1.7	1.75	1.85
	0.03	1.95	2.1	2.25	2.45	1.65	1.75	1.75	1.9
5	0.01	2.1	2.25	2.35	2.5	2.2	2.3	2.4	2.6
	0.02	2.15	2.3	2.45	2.65	2.1	2.15	2.25	2.5

(Fig. 6.140a–c) the values are taken from Table 6.50; for the key groove they are taken from Table 6.51; for the spline and thread areas of the shafts they are taken from Table 6.52. For estimation of the stress concentration at the installation points of the details with interference on the shaft, the ratios $K_\sigma/K_{d\sigma}$ and $K_\tau/K_{d\tau}$ (Table 6.53) are used.

In the case of action in the rated section of some stress concentration sources the most dangerous one is taken into account (the highest value $K_{\sigma D}$ or $K_{\tau D}$).

The stiffness calculation of the shafts is carried out in cases where their deformation (linear or angular) substantially influences the operation of the details mated with the shaft (gear wheels, bearings, and joints), producing an increase in contact stress concentration, an increase of wear, a reduction of fatigue strength, and a decrease of accuracy and evenness of rotation and displacement.

Bending and torsion stiffness are distinguished. The bending stiffness of the shafts is estimated according to the linear f and angular θ displacements under the action of forces and bending moments. Displacements are determined from the strength methods of materials. The required bending stiffness is provided under the conditions $f \leq [f]$ and $\theta \leq [\theta]$.

The allowable values $[f]$ and $[\theta]$ depend on the function of the shaft or the axles and are mainly determined by means of the correct operational conditions of

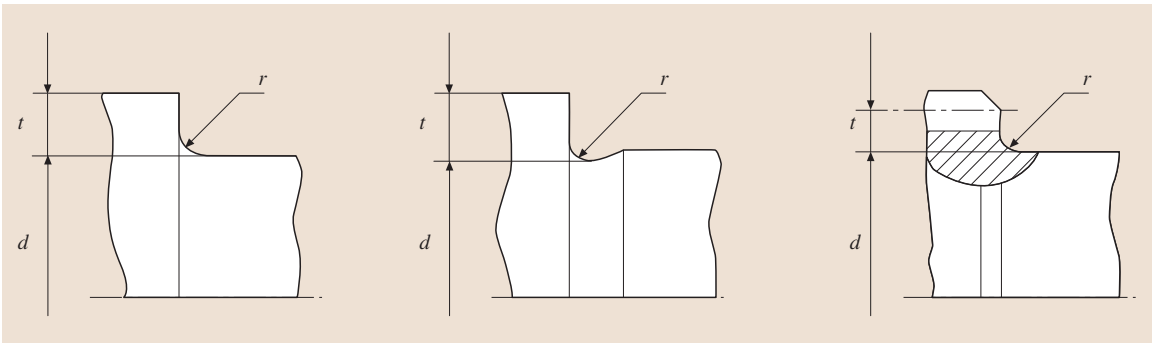


Fig. 6.140 Parameters of the step transition sections

Table 6.51 Effective stress concentration factors K_σ and K_τ for a key groove

σ_t (N/mm^2)	K_σ by groove execution with a milling cutter		K_τ
	end	disk	
500	1.8	1.5	1.4
700	2.0	1.55	1.7
900	2.2	1.7	2.05
1200	2.65	1.9	2.4

Table 6.54 Influence factors K_c of position of the gear wheels relative to the supports

Designation of the gear layout in Fig. 6.141	Coefficient K_p	Designation of the gear layout in Fig. 6.141	Coefficient K_p
1 and 2	1.2	5 and 6	0.4
3	0.8	7 and 8	0.1
4	0.6		

the gears and bearings. Elastic shaft displacements have little influence on the operation of gears with flexible couplers. In gearings they produce mutual warp of the wheels and separation of the axles, which is especially adverse for Novikov gears.

For the involute gearings of reduction gears the allowable wrap angles $[\theta]$ (radians) can be determined from the formula

$$[\theta] = 10^{-3} K_p \psi_{ba} \mathrm{HB_{me}}/600\,,$$

where K_p is a coefficient taking into account the influence of the gear wheel position relative to the bearings (Table 6.54), ψ_{ba} is a width coefficient, and $\mathrm{HB_{me}}$ is the mean hardness of the work tooth surface of the low-speed wheel.

The rigidity of the shafts rotating in the bearings must provide ease and smoothness of rotation, as well as sufficient stress distribution in contact, which finally will influence the lifetime of the bearings.

The total tolerance on the coaxiality of the cone and the outer race of the rolling bearings, which is caused by an unfavorable combination of various kinds of machining errors, assembly and deformation of the bearings, and the shaft and case details under the action of the load, is estimated from the maximum permissible angle $\theta_{\max}$ of the mutual wrap between the axles of the bearing racers that is mounted on the bearing unit.

The maximum permissible angle $\theta_{\max}$ of mutual warp of the bearing racers is defined, for which the lifetime can be proved to be not less than the required time. The values of the maximum permissible angle

Table 6.52 Effective stress concentration factors K_σ and K_τ for spline and thread sections of the shaft

σ_t (N/mm^2)	K_σ for		K_τ for splines		K_τ for thread
	Splines	Thread	Straight-sided	Involute	
500	1.45	1.8	2.25	1.43	1.35
700	1.6	2.2	2.5	1.49	1.7
900	1.7	2.45	2.65	1.55	2.1
1200	1.75	2.9	2.8	1.6	2.35

Table 6.53 Ratios $K_\sigma/K_{d\sigma}$ and $K_\tau/K_{d\tau}$ for the estimation of the stress concentration at the installation sites of the components with interference on the shaft

Shaft diameter d (mm)	$K_\sigma/K_{d\sigma}$ by σ_t (N/mm^2)				$K_\tau/K_{d\tau}$ by σ_t (N/mm^2)			
	500	700	900	1200	500	700	900	1200
30	2.6	3.3	4.0	5.1	1.5	2.0	2.4	3.05
40	2.75	3.5	4.3	5.4	1.65	2.1	2.6	3.25
50	2.9	3.7	4.5	5.7	1.75	2.2	2.7	3.4
60	3.0	3.85	4.7	5.95	1.8	2.3	2.8	3.55
70	3.1	4.0	4.85	6.15	1.85	2.4	2.9	3.7
80	3.2	4.1	4.95	6.3	1.9	2.45	3.0	3.8
90	3.3	4.2	5.1	6.45	1.95	2.5	3.05	3.9
100	3.35	4.3	5.2	6.6	2.0	2.55	3.1	3.95

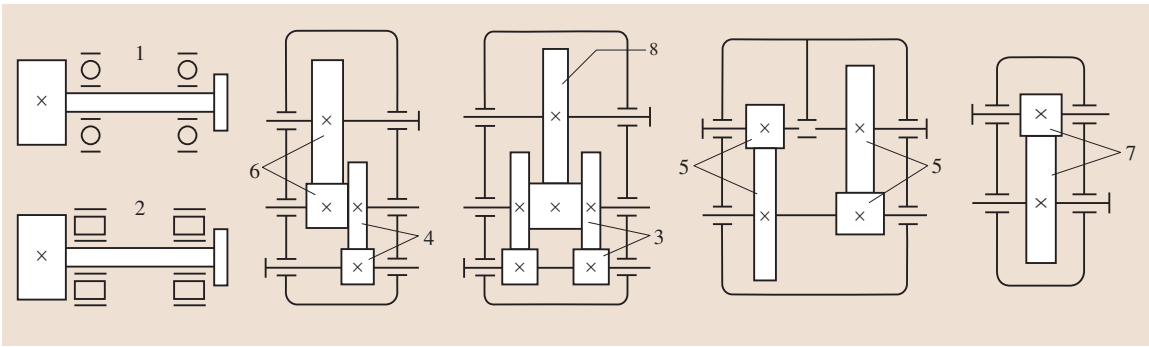


Fig. 6.141 Hookups of the gear wheels relative to the supports

are the following: $\theta_{max} = 8'$ for radial single-row ball bearings with normal radial clearance; $\theta_{max} = 5'$ for radial-stop ball single-row bearings with a contact angle of $\alpha = 26°$; $\theta_{max} = 2'$ for radial bearings with short cylindrical rollers; $\theta_{max} = 2'$ for tapered bearings with rollers without modified contact; and $\theta_{max} = 1'$ for quill axial roller bearings.

The angle θ_d of the mutual racer warp, which is caused by the deformation of the shafts and cases in the working unit, must not exceed $\theta_d = 0.2\theta_{max}$. The values θ_{max} and consequently the allowable rotary angles θ_d of the deflection shaft curve in the supports depend on the kind of bearing. The values θ_d for the self-installing bearings mounted in the rigid cases are in the range $0.4–1.6'$.

The torsion rigidity of the shafts is estimated from the twist angle under the action of the torque. Torsion rigidity is often not of vital importance and, in this case, this calculation is not carried out.

6.9.6 Shaft Design

When starting shaft (axle) design it is necessary to find out which details will be chosen, what kind of support is most appropriate for the given shaft, and under what load conditions the set of shaft details will react. It is also necessary to imagine the assembly order of the entire unit and, finally, to specify the manufacturing conditions of the shaft, i. e., the probable means of production. In the initial stage the shaft design is outlined approximately by defining the diameters of the single areas.

The input and output shafts of different machines, mechanisms, and devices that transmit torque have cylindrical or tapered end faces for mounting of half-couplings, pulleys, chain wheels, gear wheels, etc.

The permissible torques T (N m) transmitted by the shaft ends are determined from the formula

$$T = 10^{-3} K d^3 ,$$

where $K = \pi[\tau]/16(\mathrm{N/mm^2})$, and d is the diameter of the shaft end (mm).

The values of the coefficient K and their corresponding allowable twisting stresses $[\tau]$ are:

Table 6.55 Influence factor K of allowable twisting stress $[\tau]$

$[\tau]$ (N/mm^2)	K (N/mm^2)
10	2.0
14	2.8
20	4.0
28	5.6
40	8.0
56	11.2
80	16.0
112	22.4

Recommendations according to the choice of the allowable stresses $[\tau]$ and correspondingly coefficient K depending on the load conditions and mechanical characteristics of the shaft material are given in Table 6.45.

The permissible torques transmitted with the cylindrical shaft ends are given in Table 6.56; for example, the values of the allowable torques are chosen for the reduction gears and reduction gear motors as follows:

- For input shafts: $K = 8.0$ (K may chosen equal to 4.0, 5.6 or 11.2)
- For output shafts: $K = 5.6$ (K may be chosen equal to 4.0 or 8.0)

The end shaft diameter can also be chosen according to the value of the transmitted torque by the set coefficient K (N/mm^2) (Table 6.56).

Table 6.56 Allowable torques transmitted with the cylindrical ends of the shafts. The values of the torques for the shafts with a diameter of less than 6 mm are not regulated

Diameter d (mm)		Allowable torques T (N m) for the coefficient K (N/mm^2)							
1st row	2nd row	2.0	2.8	4.0	5.6	8.0	11.2	16.0	22.4
6	–	0.5	0.71	1.0	1.4	2.0	2.8	4.0	5.6
7	–	0.71	1.0	1.4	2.0	2.8	4.0	5.6	8.0
8	–	1.0	1.4	2.0	2.8	4.0	5.6	8.0	11.2
9	–	1.4	2.0	2.8	4.0	5.6	8.0	11.2	16.0
10	–	2.0	2.8	4.0	5.6	8.0	11.2	16.0	22.4
11	–	2.8	4.0	5.6	8.0	11.2	16.0	22.4	31.5
12	–	4.0	5.6	8.0	11.2	16.0	22.4	31.5	45
14	–	5.6	8.0	11.2	16.0	22.4	31.5	45.0	63.0
16	–	8.0	11.2	16.0	22.4	31.5	45.0	63.0	90.0
18	–	11.2	16.0	22.4	31.5	45.0	63.0	90.0	100
–	19	12.5	18.0	25.0	35.5	50.0	71.0	100	140
20	–	16.0	22.4	31.5	45.0	63.0	90.0	125	180
22	–	22.4	31.5	45.0	63.0	90.0	125	180	250
–	24	25.0	35.5	50.0	71.0	100	140	200	280
25	–	31.5	45.0	63.0	90.0	125	180	250	355
28	–	45.0	63.0	90.0	125	180	250	355	500
30	–	50.0	71.0	100	140	200	280	400	560
32	–	63.0	90.0	125	180	250	355	500	710
35, 36	–	90.0	125	180	250	355	500	710	1000
–	38	100	140	200	280	400	560	800	1120
40	–	125	180	250	355	500	710	1000	1400
–	42	140	200	280	400	560	800	1120	1600
45	–	180	250	355	500	710	1000	1400	2000
–	48	200	280	400	560	800	1120	1600	2240
50	–	250	355	500	710	1000	1400	2000	2800
–	53	280	400	560	800	1120	1600	2240	3150
55	56	355	500	710	1000	1400	2000	2800	4000
60	–	400	560	800	1120	1600	2240	3150	4500
63	–	500	710	1000	1400	2000	2800	4000	5600
–	65	560	800	1120	1600	2240	3150	4500	6300
70,71	–	710	1000	1400	2000	2800	4000	5600	8000
–	75	800	1120	1600	2240	3150	4500	6300	9000
80	–	1000	1400	2000	2800	4000	5600	8000	11 200
–	85	1120	1600	2240	3150	4500	6300	9000	12 500
90	–	1400	2000	2800	4000	5600	8000	11 200	16 000
–	95	1600	2240	3150	4500	6300	9000	12 500	18 000
100	–	2000	2800	4000	5600	8000	11 200	16 000	22 400
–	105	2500	3150	4500	6300	9000	12 500	18 000	25 000
110	–	2800	4000	5600	8000	11 200	16 000	22 400	31 500
–	120	3150	4500	6300	9000	12 500	18 000	25 000	35 500
125	–	4000	5600	8000	11 200	16 000	22 400	31 500	45 000
–	130	4500	6300	9000	12 500	18 000	25 000	35 500	50 000
140	–	5600	8000	11 200	16 000	22 400	31 500	45 000	63 000
–	150	6300	9000	12 500	18 000	25 000	35 500	50 000	71 000
160	–	8000	11 200	16 000	22 400	31 500	45 000	63 000	90 000
–	170	9000	12 500	18 000	25 000	35 500	50 000	71 000	100 000
180	–	11 200	16 000	22 400	31 500	45 000	63 000	90 000	125 000

The diameters of the other areas are fixed in shaft structure development, taking into account their functionality, production technique, and assembly. To this end empirical dependencies can be used. Thus the diameters of the crankshaft journals are determined from empirical formulas depending on the diameter of the motor cylinder; the diameters of the work spindles are calculated depending on the main geometry of the machine, etc.

Before adjustment of the shaft structure important questions must be solved such as the transmission method of the torque in the joint shaft–hub and the fastening manner of the details onto the shaft from the axial displacement.

The shaft strength with the action of the varying stresses is to a great extent caused by its structural forms at the transition points between the grades, positions of splines, grooves, openings, turnings, etc., where there is stress concentration of bending and twist. To increase the fatigue strength of the shafts, various techniques of technological strengthening and structural improvement that provide stress concentration reduction are applied.

Thus, for example, if the transition from the shaft area with the diameter d to the area of the major diameter is carried out using hollow chamfers, the radius r_1 (Fig. 6.142a) of the hollow chamfer should be as large as possible, as the stress concentration factor increases with the reduction of the ratio r_1/d. For $r_1/d < 0.1$ the stress concentration factor can be equal to 2 or more. For the purpose of decreasing the stress concentration, elliptic hollow chamfers with dimensions $b = (0.4–0.45)d$ and $a = 0.4b$ (Fig. 6.142b), or hollow chamfers outlined with two interfacial circular arcs are used.

In those cases when, due to some structural consideration, small radii of the hollow chamfers have to be used, it is recommended that a recessed undercut of the shaft into the shoulder be carried out (Fig. 6.142c).

Balancing grooves on the shaft (Fig. 6.142d) and on the end of the mating component (Fig. 6.142e) reduce the stress concentration on the shaft surface from the component fit with interference. These measures decrease the stress concentration by 15–25%. Inclusion of the circular turnings on the ends increases the yielding of the hub, making the pressure distribution more even along the length of the joint. Extension by ~ 5% of the shaft area diameter at the point of interference connection increases the shaft yielding under the hub ends, thus reducing the stress concentration factor. A key groove executed with a disk mill (Fig. 6.142f) generates less stress concentration than that machined with a shank milling cutter. Involute splines generate less stress concentration in comparison with straight-sided splines.

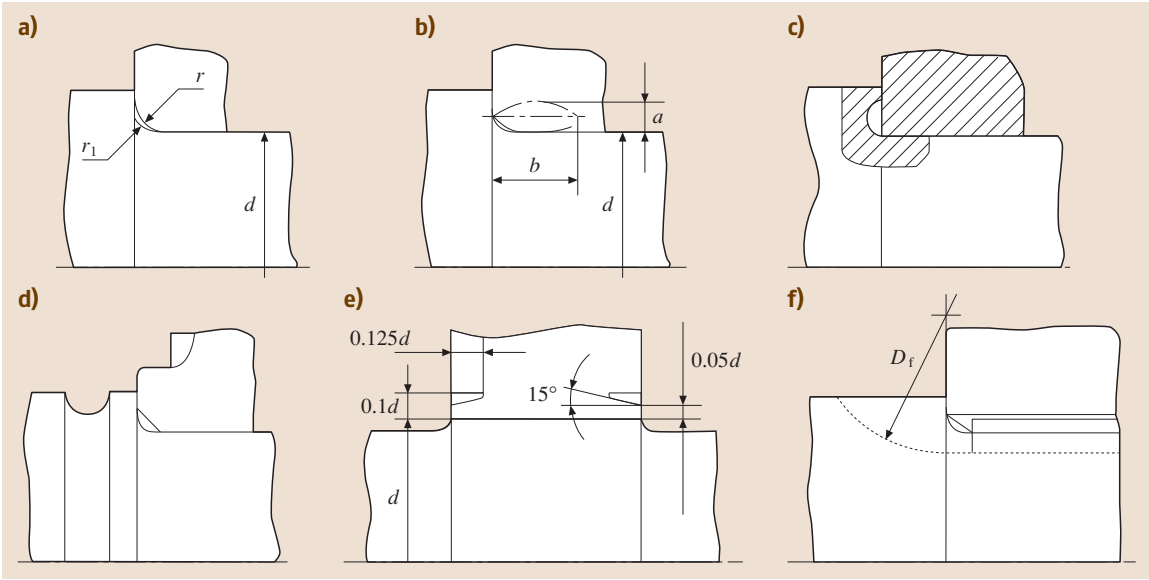


Fig. 6.142a–f Increase of the fatigue strength of the shafts using **(a)** a hollow chamfer of constant radius, **(b)** an an elliptic hollow chamfer, **(c)** a hollow chamfer with recessed undercut, **(d)** an unloading groove on the shaft and **(e)** in the mating component, **(f)** a key groove with a disk cutter, but not an end-milling cutter

The spline connection reduces the fatigue strength of the shaft less than the key joint.

Component fastening on the shaft using lock screws, adjusting nuts, cutting rings, etc., increases stress concentration and consequently reduces the fatigue strength of the shaft. Thus it is advisable to use axial fastening for the components. When an opening application for the lock screws or pins, nut threads, grooves for the elastic rings, etc., cannot be avoided, measures should be adopted to decrease stress concentration at these points.

Shaft hardening by structural means at the positions of the transverse openings can be carried out by the following methods: countersinking the hole, removal of a flat along the hole, and insertion of a bronze (a material with a lower coefficient of elasticity) bushing into the hole. These measures can decrease the stress concentration by 20–40% or more.

The thread is characterized by considerable stress concentration. The stress concentration factor for the thread substantially depends on the thread vee-radius R between the threads. For high-duty shafts it is recommended to use a thread with vee-radius $R = (0.125–0.144)P$, where P is the thread pitch.

Another reason for stress concentration is fretting wear (friction corrosion) that results from the slightly varying relative displacements of the shaft and the mated component, which are in turn caused by flexural or torsional strain. The stress concentration is especially strong in those cases in which the component is set on the shaft with interference and when it transmits loads to the shaft.

The fatigue strength of the shafts under the hubs can be raised with plastic forming (breaking-in with a roller), chemicothermal treatment (nitriding), surface hardening, and treatment with a laser beam and plasma. The quality of the surface layer in the weak sections distinctly affects the fatigue strength of the shaft, especially at stress concentration sites. The structure of the journals is caused by the type of shaft bearing applied (rolling bearing or sleeve bearing). The journal diameter of the friction bearing is chosen depending on the required strength and rigidity of the shaft and the overall dimensions of the whole structure. To increase the reliability of the friction bearing it is usually helpful to increase the journal diameter, but it must be borne in mind that journals are end faces of the shaft and according to the assembly conditions they are designed to have a smaller diameter than the middle parts of the shaft.

To decrease wear the journals are heat treated or chemicothermally treated (hard-surfacing, cementing, nitriding), leaving the core viscous. Above all this is relevant for bearing units working with contaminated lubricants, for which abrasion is typical.

If the shaft has rolling bearings, the diameter and length of the shaft journals for the bearing are determined from the dimensions of the chosen bearing. Technical requirements (roughness, deviation of the form and position) for the mounting and bearing front faces must meet the requirements for ball and roller bearings.

Tolerance ranges for the diameters of the mounting shaft surfaces, as well as the fit for the bearing joint with the shafts (axles), are fixed according to the accuracy grades of the bearings. The nature of the mating of the bearing with the shaft and the choice of the fit depend on whether the inner race of the bearing rotates or not relative to the radial load that affects it, and on the direction and value of the acting load intensity, etc.

When fastening the inner races of the rolling bearings on the shaft in the axial direction structural measures must be taken to provide the correct mounting, dismantling, and required maintenance of the bearings in operation.

As a result of the temperature increase during operation of the product, the shaft can become elongated, which is why fixing of the shaft from the axial displacement must be done such that the shaft elongation does not cause jamming of the bearings or lead to the occurrence of secondary stress.

The method of shaft fastening from the axial displacement is chosen depending on the kind of bearings that are mounted on the shaft (adjustable or non-adjustable), and on the working conditions of other components that mate with the shaft.

The machining accuracy of the journals (necks) for friction and rolling bearings is of great consequence. Because of the different component types that are set on the shafts and the axles, the greatest demands are made on the mounting of gear and worm wheels and pulleys of high-speed belt drives with regard to the coaxiality of the shaft sections bearing these components relative to the journals (necks). For gears with toothing this results from the necessity of providing the standards of kinematic accuracy and of contact; for pulleys it is necessary to decrease imbalance and, consequently, dynamic loads and vibrations.

Thus, for example, the coaxiality tolerance of a mounting shaft surface with a diameter of 56 mm for gear wheels, a pitch diameter of 240 mm with kinematic accuracy degree 7 of the gear is 0.025 mm.

Certain accuracy demands, if necessary, can also be made for other sections or structural shaft components;

for example, the tolerances on the symmetry and parallelism of the key groove of the shaft axles are fixed in order to provide the possibility of assembling the shaft with the component mounted on it and to provide even contact of the key and shaft surfaces.

6.9.7 Drafting of the Shaft Working Drawing

Introduction

Dimensioning. In the working drawings the *minimum* number of dimensions must be set, but they must be *sufficient* for the production and control of the component. The dimensions given in the drawings can be classified as being:

- *Functional*, determining qualitative product indexes: dimensions of the assembly measuring chains, mating dimensions, diameters of the shaft sites for gear and worm wheels, couplings, bearings and other components, and thread dimensions on the shafts of the adjusting nuts
- *Free* (dimensions of the nonjoining surfaces)
- *Reference*

Functional dimensions are set in the working drawings of the components, having been taken from the drawing of the assembly unit (reduction gear, gearbox) and from the layouts of the dimensional chains. Free dimensions are set, taking into account the fabrication technique and control convenience. Reference dimensions are not subjected to execution according to the given working drawings and are not controlled during component manufacture. Reference dimensions are marked with an asterisk and a notation such as "* Dimensions for reference" is added in the standards.

Extreme Dimensional Deviations

For all the dimensions given in the working drawings extreme deviations are indicated in millimeters. It is permissible not to indicate extreme deviations of dimensions that fix areas of different roughness and accuracy of the same surface, of the heat-treated zone, the coating and knurling zone, as well as the diameters of the knurled surfaces. In these cases, the sign "≈" is marked directly on such dimensions. If necessary extreme deviations of the *rough* or *very rough* accuracy degree according to the Russian standard [6.55] (Table 6.57) are set for these dimensions instead of using this sign.

If extreme deviations (tolerances) are not given individually for the appropriate nominal dimensions, the *overall dimensional tolerances* according to the Russian standard [6.55] are applied, fixed according to four accuracy degrees: accurate f, mean m, rough c, and very rough v (Table 6.57). For the choice of the accuracy degree the common accuracy of the corresponding industry is taken into account.

The overall dimensions are applied for the following dimensions with undisclosed individually extreme deviations:

- For linear dimensions (e.g., outer and inner diameters, radii, distances, shoulder dimensions, dimensions of the dull edges, outer rounded radii and chamfer dimensions)
- For angular dimensions, including angular dimensions that are usually undisclosed, i. e., right angles or angles of regular polygons
- For linear and angular dimensions, which are obtained by ready-mounted component machining

References to the overall tolerances of the linear and angular dimensions are given in the standards, indicating the number of the standard and the letter symbol of the accuracy degree required, e.g., for the accuracy degree *mean*: "Overall tolerances according to GOST 30893.1-m" or "GOST 30893.1-m".

The individual extreme deviation of the linear dimensions is indicated according to one of the three following methods:

- Reference designations of the tolerance ranges, e.g., 63H7
- Values of the extreme deviations, e.g., $64^{+0.030}$

Table 6.57 Extreme deviations of the linear dimensions according to [6.55]

Accuracy degree	Extreme deviations for the intervals of the dimensions (mm)					
	0.5–3	> 3–6	> 6–30	> 30–120	> 120–400	> 400–1000
Accurate f	±0.05	±0.05	±0.1	±0.15	±0.2	±0.3
Mean m	±0.10	±0.10	±0.2	±0.30	±0.5	±0.8
Rough c	±0.20	±0.30	±0.5	±0.80	±1.2	±2.0
Very rough v	–	±0.50	±1.0	±1.50	±2.5	±4.0

- Reference designation of the tolerance ranges with indication of the extreme deviation values in brackets to the right: $18P8\left(^{-0,018}_{-0,045}\right)$

The first method is recommended in the case of nominal dimensions, which are included in the series of standard numbers [6.83]. The second method is used in the case of nonstandard numbers on the nominal dimensions, and the third is used with standard numbers, but with inadvisable tolerance ranges.

Extreme deviations of chain dimensions are assigned according to the results of the probability-theoretical calculation of the corresponding dimensional chains. Approximately extreme deviation of the chain dimensions can be taken according to the compensation method:

- If compensator is a component that is scraped or ground according to the results of the measurement by assembly, with a view to decreasing the machining allowance of the tolerance ranges of the chain dimensions should be assumed: of the openings H9, of the shafts h9, others $\pm$ IT9/2.
- If a gasket package serves as a compensator, the tolerance ranges of the chain dimensions are assumed to be H11, h11, $\pm$ IT11/2.
- If a thread pair serves as a compensator, as a consequence of its wide compensating possibilities, the tolerance dimensional ranges are assumed to be: H14, h14, $\pm$IT14/2.

Extreme deviations of the thread diameters are shown in the component working drawings in accordance with the fits of the threaded connections that are given in the working drawings of the assembly units, for example, for the threads in the openings M20-7H, M16-3H6H, M30 ×1.5-2H5C, and for the threads on the shafts M42-8g, M16-2m, M30 ×1.5-2r.

Form Tolerances and Tolerances on the Surface Position

During machining of the components errors arise not only in the linear dimensions, but also in the geometry, as well as the errors in the relative position of the axles, surfaces, and structural components of the details. These errors can exert an unfavorable influence on the efficiency of the machinery, producing vibrations, dynamic loads, and noise. The first group of accuracy requirements is caused by the installation of the rolling bearings (Russian standard [6.89]). It is important for rolling bearings that the rolling paths of the racer are not distorted. Racers are very compliant and on installation they adopt the form of the mounting surfaces of the shafts and cases. To decrease the shape defects of the rolling paths form, tolerances are set for the mounting surfaces of shafts and cases.

The relative warp of the outer and inner races of the bearings increases shaft rotation and power waste resistance, and reduces the lifetime of bearings. Race warp can be caused by:

- Axial deviations of the mounting surfaces of the shafts and the case
- Perpendicularity deviations of the datum faces of the shaft and case
- Deformations of the shaft and case in the working unit

To limit these deviations the tolerances on the mounting surface position of the shaft and case are set in the working drawings.

The second group of accuracy requirements results from the necessity to abide by kinematic accuracy standards and contact standards of tooth and worm gears [6.28, 29, 41]. The achievement of the requisite accuracy depends on the positional accuracy of the mounting surfaces and the datum faces of the shafts, as well as the mounting openings and the datum faces of the wheels. Thus, the tolerances on the datum face position are set in the working drawings of the shafts, gear, and worm wheels.

The third group of accuracy requirements is caused by the need for limitation of possible component unbalance. Allowable imbalance values are defined in [6.86] depending on the kind of product and its operating conditions. The standards of allowable imbalance are described by the equation $en = \text{const.}$, where e is a specific imbalance (g mm/kg), which is numerically equal to the displacement of the mass center from the rotation axles (micrometer), and n is a rotational frequency (min^{-1}). In this respect it is convenient to make *demands* on the single component surfaces *in the form of the coaxiality tolerances* in the working drawings.

Base axles and surfaces are indicated in the working drawings with equilateral hatched triangles connected with a frame, where the designation of the base is written with a capital letter.

If the tolerance on the form or the position is not given individually for the appropriate element of the detail, overall tolerances on the form and position according to [6.56] are applied, being fixed for three accuracy degrees (in decreasing accuracy order): H, K, and L. By the choice of the accuracy degree, the com-

Table 6.58 Standard values form and position tolerances

1	1.2	1.6	2	2.5	3	4	5	6	8
10	12	16	20	25	30	40	50	60	80
100	120	160	200	250	300	400	500	600	800

mon accuracy of the appropriate industry is taken into account.

References to the overall tolerances on the form and position are given in technical specifications, indicating the number of the standard and the letter symbol of the accuracy degree, e.g., for the accuracy degree *K*: "Overall form and position tolerances according to GOST 30893.2-K" or "GOST 30893.2-K."

The overall tolerances on the dimensions, form, and position are given in the technical specifications with a note like: "Overall tolerances according to GOST 30893.2-mK" or "GOST 30893.2-mK."

In the cited example *m* is the *mean* accuracy degree of the overall tolerances for the linear dimensions according to GOST 30893.1, and *K* is the accuracy degree of the overall tolerances on the form and position according to GOST 30893.2.

The form and position tolerances are rounded to the nearest number (μm) from the series of the preferable numbers after their determination (Table 6.58).

Surface Roughness

Of the roughness parameters set in [6.82] the following are the most widely used in machine construction:

- *Ra* is an arithmetic mean profile deviation (micrometers) (the main of the high-altitude roughness parameters is set for all unmachined surfaces)
- *Rz* is a cusp height of the profile (micrometers) (determined according to five measurements of the cusp heights, set for the surfaces after casting, forging, and engraving)
- *tp* is a relative bearing length of the profile, where *p* is a value of the profile section level

The parameter *tp* contains the contact area estimation of the mating surfaces. It is set for mating surfaces that require tightness, contact rigidity, wear resistance, or cohesive resistance (e.g., components connected by interference).

For the designation of surface roughness special characters are used in the working drawings according to [6.82]. Designation of the dominant roughness is shown in the upper-right corner of the working drawing field.

The values of the roughness parameter *Ra* can be taken from Table 6.59.

Positioning of Dimensions, Datum Designations, Form and Position Tolerances, Roughness, and Technical Requirements in the Component Drawing

For convenience, all the information needed for the component manufacture is organized into the following system. In the component working drawings, fir solids of rotation (shafts, pinion-shafts, worms, wheels, cartridges, and bearing caps) the following should be set (Fig. 6.143):

- Axial linear dimensions: located under the component drawing, in as small as possible number of levels of dimension lines (2–3).
- Datum reference designations: under the component drawing.
- Reference designation of the form and position tolerances: under the component drawing in one or two levels.
- Reference designation of the roughness parameters: in the upper parts of the component drawing. In the faces: under the component drawing. In both cases, reference roughness designations are located in the immediate proximity to the dimension line.
- Specify line notes, indicating surfaces for heat treatment and coating: above the component drawing.

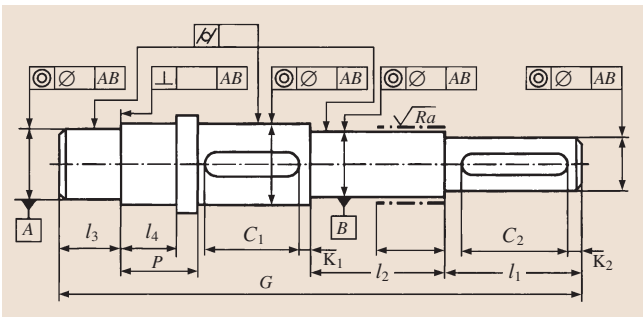


Fig. 6.143 Location of the dimensions, base designations, form tolerances, and position tolerances on the component drawing

Table 6.59 Recommended values of roughness *Ra*

Type of surface	*Ra* (μm)
Mounting surfaces of the shafts and the cases from steel for the rolling bearings of the *normal* accuracy degree for:	
d or *D* up to 80 mm	1.25
d or *D* over 80 mm	2.5
Mounting surfaces of the cases from iron for the rolling bearings of the *normal* accuracy degree for:	
D up to 80 mm	2.5
D over 80 mm	3.2
Pin shoulder faces of the shafts and the cases for stationing of the rolling bearings of the *normal* accuracy degree	2.5
Shaft surfaces for interference joints	0.8
Pin shoulder faces for positioning of the gear, worm wheels with the ratio of the opening hub length to its diameter:	
$l/d < 0.7$	1.6
$l/d \geq 0.7$	3.2
Shaft surfaces for cup-type seals	0.32
Case (cover) surfaces for cup-type seals	1.6
Grooves, bevels, hollow chamfer radii on the shafts	6.3
Surfaces of the key grooves on the shafts:	
effective	3.2
noneffective	6.3
Surfaces of the key grooves in the openings of the wheels, pulleys:	
effective	1.6
noneffective	3.2
Spline surfaces on the shafts:	
– tooth surface of the joint:	
fixed	1.6
sliding	0.8
– cylindrical surfaces, centering joining:	
fixed	0.8
sliding	0.4
– cylindrical surfaces, noncentering joining	3.2
Spline surfaces in the openings of the wheels, pulleys, chainwheels:	
– tooth surface of the joint:	
fixed	1.6
sliding	0.8
– cylindrical surfaces, centering joining:	
fixed	1.6
sliding	0.8
– cylindrical surfaces, noncentering joining	3.2

Table 6.59 (cont.)

Type of surface	*Ra* (μm)
Opening surfaces of the hubs with interference connections	1.6
Hub faces of gear, worm wheels positioned along the pin shoulder face of the shaft with the ratio of the opening length to its diameter	
$l/d < 0.7$	1.6
$l/d \geq 0.7$	3.2
Hub faces of gear, worm wheels, along which the rolling bearings of the accuracy degree *normal* are positioned	1.6
Free (noneffective) faces of gear, worm wheels	6.3
Working tooth surfaces of gear wheels with external toothing:	
With the module ≤ 5 mm	1.25
With the module > 5 mm	2.5
Working surfaces of the worm coils;	
cylindrical	0.63
concave	1.25
Working tooth surfaces of worm wheels	1.6
Cusp surfaces of the wheel teeth, worm coils, chain wheel teeth	6.3
Bevels and recesses on the wheels	6.3
Opening surfaces in the covers for the rubber glands	1.6
Working surface of the belt pulleys	2.5
Working tooth surface of the chainwheels	3.2
Opening surfaces for bolts, screws, stud-bolts	12.5
Bearing surfaces for bolt, screw, and nut heads	6.3

Technical requirements are located above the main inscription, and if there is not enough space they are placed to the left of the main inscription. Technical requirements are written in the following order:

1. Requirements for the material, workpiece, heat treatment, and the material properties of the finished part (HB, HRC)
2. Guidelines about dimensions (dimensions for references, rounded radii, angles, etc.)
3. Overall tolerances on the dimensions, forms, and positions
4. Tolerances on the forms and mutual surface position, for which there are no conventional graphic characters in [6.80]
5. Surface quality requirements (guidelines about finish, coating, roughness)
6. Units of measurement that have to be indicated for the dimensions and extreme deviations given in the technical requirements

Performance of the Shaft Working Drawing

Dimensions and Extreme Deviations. In the shaft working drawings the mating, chain, and overall and free dimensions are set. Figure 6.143 shows a method for axial dimensioning of the shaft. The dimensions are indicated in this figure: C_1 and C_2 are the matings (lengths of the key grooves); G and P are overall and chain dimensions, K_1 and K_2 coordinate the position of the key grooves, which is convenient for the control with a vernier caliper or with a trammel; l_1 is the length of the shaft extension (conjunctive dimension), l_2 and l_3 are the lengths of the mating surfaces. The dimensions l_1, l_2, l_3, and l_4 correspond to the consecutive phases of the shaft turning. In this example, the dimen-

sions C_1, C_2, and P are functional, whereas the others are free.

In the shaft working drawings by means of the note on the magnification scale (4 : 1) the form and dimensions of the grooves are given for the outlet of the grinding wheel and the cavities for the outlet of the thread-cutting tool. The depth of the key groove, the dimension t_1 according to the Russian standard [6.90] (Fig. 6.144a), is also set. If the key groove located on the shaft end is through, it is convenient for the control to the dimension $(d - t_1)$.

On the conic shaft end the depth of the key groove t_1' (Fig. 6.144b) is determined from the formula

$$t_1' = t_1 + 0.025l\,,$$

where t_1 is assumed for the shaft diameter d_2 according to [6.85].

Sometimes the depth t_1 of the key groove is shown on the mean diameter d_2 of the shaft. In this case, the distance to the measuring section is set in the shaft drawing (Fig. 6.144b).

The tolerance ranges are specified for the mating dimensions in accordance with the fits shown in the component drawing. Extreme deviations in the tolerance ranges for the recommendations given above are assigned for the chain dimensions. Extreme deviations, mostly of the *mean accuracy degree* (Table 6.57), are specified for the free dimensions. Designation of the tolerance range is indicated for the width of the key groove: for the straight key P9, and for the semicircular key N9.

Extreme deviation of the key groove depth t_1 (Fig. 6.144) with key section up to 6×6 mm: +0.1 mm; 6×6 – 32×18 mm + 0.2 mm. Extreme deviations for the dimensions $(d - t_1)$ are assigned accordingly: −0.1 and −0.2 mm.

The tooth length of the total profile to the runout is given in the shaft working drawings that have the elements of the spline connections. For the roughness designation on the lateral surfaces the profile of one tooth is shown. The reference designation of the spline connection elements of the shaft according to the corresponding standard is given on the line-note in the technical requirements.

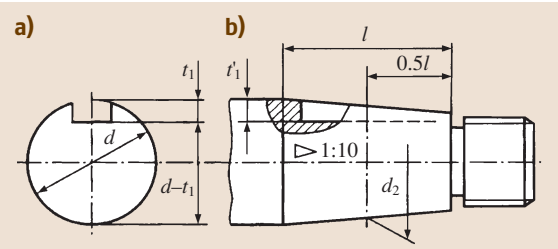


Fig. 6.144a,b Definition of the key groove depth on the end of **(a)** the cylindrical shaft and **(b)** the tapered shaft

Form Tolerances and Surface Position Tolerances

The shaft in the working unit rotates in the rolling bearings. As the rolling bearings are manufactured with relatively high accuracy, the production errors of their components are usually neglected. This is why the work axle is *a general axle*, designated with the letters AB in Fig. 6.145. The general axle is a straight line passing through the cross-points of each of the two mounting surfaces for the rolling bearings with the mean cross-sections of these surfaces.

As a consequence of unavoidable errors, the general axle does not coincide with the rotation axles of the shaft after manufacture. The required accuracy demands of single-component production are assigned in the shaft drawing. The guidelines for the value determination of the form and surface position tolerances are given in Table 6.60 in accordance with the positions in Fig. 6.145.

The destination of each of the form or position tolerances is as follows:

- The tolerance of the cylindrical mounting surfaces for the rolling bearings (position 1) is specified to limit geometry deviations of these surfaces and in this way to limit geometry deviations of the roller paths of the racers (according to [6.89] the following individual constituents of this tolerance should

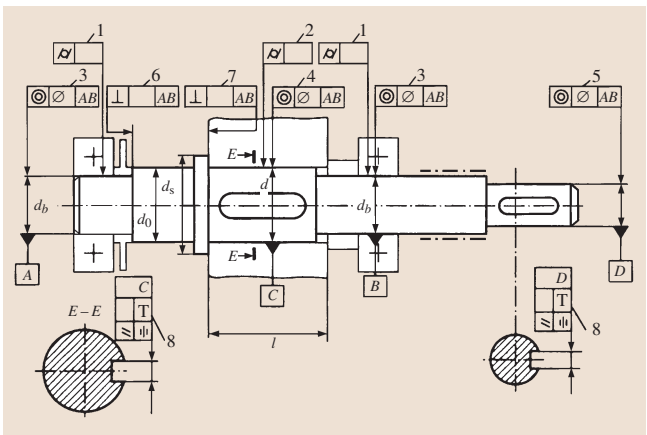


Fig. 6.145 Designation of the production accuracy requirements of single shaft elements

Table 6.60 Recommendations concerning the determination of the form tolerances and position tolerances of the shaft surfaces

Position in Fig. 6.145	Tolerance
1, 2	$T_{ci} \approx 0.5t$, where t is a surface dimension tolerance
3	T_{so} according to Table 6.61 depending on the bearing type
4	T_{so} on the diameter d according to Table 6.62. The tolerance accuracy degree is according to Table 6.63
5	$T_{so} \approx 60/n$ for $n > 1000$ min^{-1}; tolerance is in mm
6	T_{pr} on the diameter d_0 according to Table 6.64. Tolerance accuracy degree by bearing positioning: ball bearings – 8, roller bearings – 7
7	T_{pr} on the diameter d_s by $l/d < 0.7$ according to Table 6.64. Tolerance accuracy degree is according to Table 6.65
8	$T_{pa} \approx 0.5t_{sp}$; $T_{si} \approx 4t_{sp}$, where t_{sp} is a width tolerance of the key groove

Table 6.61 Tolerances of coaxiality T_{sow} and T_{sok} for mounting surfaces of the shaft and the case in bearing units. T_{sow} and T_{sok} are coaxiality tolerances of the mounting surface of the shaft and the case with length $B = 10$ mm in diametral form. For length B_1 of the slot the tabulated value should be multiplied by $0.1B_1$. θ is the allowable angle of mutual warp of the racers, caused by deformations of the shaft and the case in the working unit

Bearing type	T_{sow} (μm)	T_{sok} (μm)	θ (angle min)
Radial ball, single-row	4	8	1.6
Radial-thrust ball, single-row	3	6	1.2
Radial with short cylindrical rollers:			
without modified contact	1	2	0.4
with modified contact	3	6	1.2
Taper roller:			
without modified contact	1	2	0.4
with modified contact	2	4	0.8
Needle roller single-row			
without modified contact	0.5	1	0.2
with modified contact	2	4	0.8
Radial ball and roller double-row spherical	6	12	2.4

Table 6.62 Coaxiality tolerances according to [6.88]

Dimension range (mm)	Coaxiality tolerance (μm) for tolerance accuracy degree:				
	5	6	7	8	9
over 18 up to 30	10	16	25	40	60
Over 30 to 50	12	20	30	50	80
Over 50 to 120	16	25	40	60	100
Over 120 to 250	20	30	50	80	120
Over 250 to 400	25	40	60	100	160

be controlled: roundness accuracy tolerance, tolerance of the longitudinal section profile, diameter variability tolerance in the cross and longitudinal section).

- The tolerance on cylindrical shape (position 2) of the mounting shaft surfaces is set in their installation sites with interference of gear and worm wheels to limit pressure concentration.

Table 6.63 Recommended accuracy degrees of coaxiality tolerance. The accuracy degree of the coaxiality tolerances of the slots are for the wheels of tooth (numerator) and worm (denominator) gears

Kinematic accuracy degree of the gear	Accuracy degree of the coaxiality tolerance with the diameter of the pitch circle (mm)		
	over 50 up to 125	over 125 up to 280	over 280 up to 560
6	5/6	5/6	6/7
7	6/7	6/7	7/8
8	7/8	7/8	8/9
9	7/8	8/9	8/9

Table 6.64 Tolerances of parallelism and perpendicularity in compliance with GOST 24643-81

Dimension range (mm)	Parallelism, perpendicularity tolerance (μm) for tolerance accuracy degree:					
	5	6	7	8	9	10
Over 16 up to 25	4	6	10	16	25	40
Over 25 to 40	5	8	12	20	30	50
Over 40 to 63	6	10	16	25	40	60
Over 63 to 100	8	12	20	30	50	80
Over 100 to 160	10	16	25	40	60	100
Over 160 to 250	12	20	30	50	80	120
Over 250 to 400	16	25	40	60	100	160

Table 6.65 Recommended accuracy degrees of perpendicularity tolerance

Wheel type	Accuracy degree of the perpendicularity tolerance by accuracy degree of the gear according to the contact standards		
	6	7 and 8	9
Gear wheels	5	6	7
Worm wheels	6	7	8

- The coaxiality tolerance of the mounting surfaces for rolling bearings relatively to their mutual axles (position 3) is set to limit the warp of the rolling bearing racers.
- The coaxiality tolerance of the mounting surface for the gear and worm wheel (position 4) is specified to guarantee kinematic accuracy standards and contact standards of tooth and worm gears.
- The coaxiality tolerance of the mounting surface for half-coupling, pulley, chainwheel (position 5) is set to decrease the imbalance of the shaft and the components installed on this surface. The coaxiality tolerance according to position 5 is set by a rotational frequency of more than $1000\,\mathrm{min}^{-1}$.
- The perpendicularity tolerance of the datum shaft face (position 6) is specified to decrease the warp of the racers and geometry distortion of the rolling path of the inner race.
- The perpendicularity tolerance of the datum shaft face (position 7) is set only when mounting narrow gear wheels ($l/d < 0.7$) on the shaft. The tolerance is set to guarantee execution of the contact standards of the teeth in the gear.
- Symmetry and parallelism tolerances of the key groove (position 8) are specified to guarantee the possibility of shaft assembly with the component installed on it and an even contact surface between the key and the shaft.

The tables referred to in Table 6.60 are given below. The values of θ according to Table 6.61 are used by shaft rigidity checking. Figure 6.146 shows an example of a shaft drawing.

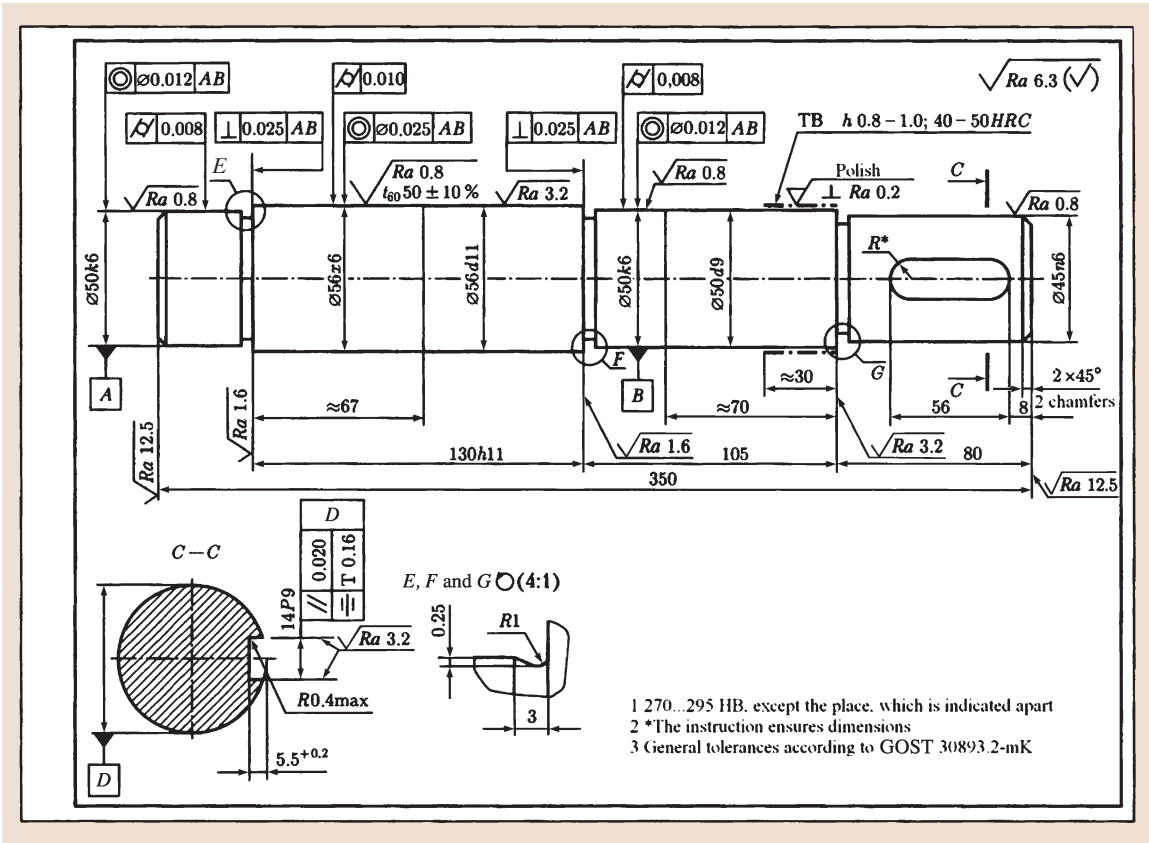


Fig. 6.146 Execution example of a working shaft drawing

Example: Determination of Form and Position Tolerances of Mounting Shaft Surfaces (Fig. 6.146)

Tolerances on cylindrical shape (positions 1 and 2, Table 6.60)

surface ⌀ $50k6 - t = 16\,\mu m$ ([6.53]),

$T_{ci} = 0.5 \times 16 = 8\,\mu m$, $T_{ci} = 0.008\,mm$;

surface ⌀ $56x6 - t = 19\,\mu m$,

$T_{ci} = 0.5 \times 19 = 9.5\,\mu m$, $T_{ci} = 0.010\,mm$.

Coaxiality tolerances (position 3) of the mounting surfaces ⌀50*k*6 with length $B_1 = 35\,mm$. For a ball radial bearing according to Table 6.61 $T_{so} = 0.1 B_1 T_{sow} = 0.1 \times 35 \times 4 = 14\,\mu m$. After round-up this yields $T_{so} = 0.012\,\mu m$.

The coaxiality tolerance (position 4) of the mounting surface ⌀56×6. With the kinematic accuracy degree 7 of the gear for the gear wheel with pitch diameter 240 mm according to Table 6.63, the accuracy degree of the coaxiality tolerance is 6. According to Table 6.62 therefore $T_{so} = 0.025\,mm$.

The rotational shaft frequency $n < 1000\,min^{-1}$ and tolerance in position 5 are not set. The perpendicularity tolerance (position 6) of the shaft collar with diameter $d_0 = 56\,mm$. For a ball radial bearing the tolerance accuracy degree is 8 (Table 6.60). According to Table 6.64 the tolerance is $T_{pr} = 0.025\,mm$.

The tolerance in position 7 is not set, because there is no collar on the shaft.

Parallelism and symmetry tolerances of the key groove (position 8): dimensional groove tolerance [6.53]: $t_{sp} = 43\,\mu m$. Then $T_{pa} = 0.5 t_{sp} = 0.5 \times 43 = 21.5\,\mu m$; $T_{pa} = 0.020\,mm$. $T_{si} = 4 t_{sp} = 4 \times 43 = 172\,\mu m$; $T_{si} = 0.16\,mm$.

6.10 Shaft–Hub Connections

6.10.1 Key Joints

For transmission of torque between the shaft and the gear, worm wheels, pulleys, chainwheels, half-couplings, etc., mounted on it, *straight and semicircular* keys are mostly used. Stationary key joints are the most efficient, which combine with an interference fit of the hub on the shaft, provide with hub centering on the shaft and exclude contact corrosion. *Feather and sliding keys* are sometimes used in movable joints of the hub with the shaft in the axial direction (e.g. a movable pinion unit of the speed gearbox). Because of their low load-carrying capacity, these joints are changed into sliding spline connections in the case of a new design [6.76, 90–99].

Straight keys have a rectangular cross-section, and their ends are chamfered from one or two faces (Fig. 6.147a) or flat (Fig. 6.147b). Rounded faces of the key ease installation of the component on the shaft in the case of a slight mismatch of the lateral faces of the key and the groove in the hub. The standard foresees definite values of the width b and the height h of the key cross-section and the groove depth on the shaft t_1 and in the hub t_2. The lengths l of the keys are also standardized. Thus, the key represents a steel bar mounted in the grooves of the shaft and the hub.

The groove in the hub is made by means of a broaching or shaping cutter. The groove for the key on the shaft is made with a shank milling or disk cutter. Manual fitting is often needed for grooves made with a shank-milling cutter. Side narrower key faces with the height h are active.

The main efficiency criterion of the key joints is strength. The keys are chosen from standard tables depending on the shaft diameter, and then the joint is checked for strength. The dimensions of the keys and the grooves are matched so that their shearing and bending strength is guaranteed if the condition of bearing strength is met, which is why the main analysis of the key joints is a bearing calculation.

The operating mode, the material strength of the components, and fit type are taken into consideration by the choice of the allowable stresses $[\sigma]_{st}$.

Joints with straight keys (Fig. 6.148) are checked for bearing according to the strength condition

$$\sigma_{st} = 2\times10^3 T/(dkl_c) \le [\sigma]_{st} ,$$

where T is a torque (N m), d is the shaft diameter (mm), $k = h - t_1$ is an extension of the key from the shaft (the entry depth of the key into the hub) (mm), l_c is a rated key length (mm) (see Fig. 6.147), and $[\sigma]_{st}$ is an allowable bearing stress (N/mm^2).

By the projection calculation the rated key length is determined from the strength condition

$$l_c \ge 2\times10^3 T/(dk[\sigma]_{st}) .$$

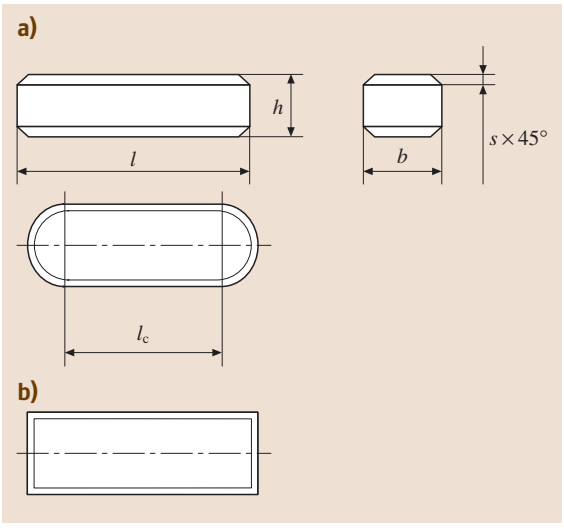


Fig. 6.147a,b Straight keys. **(a)** With chamfered faces, **(b)** with flat faces

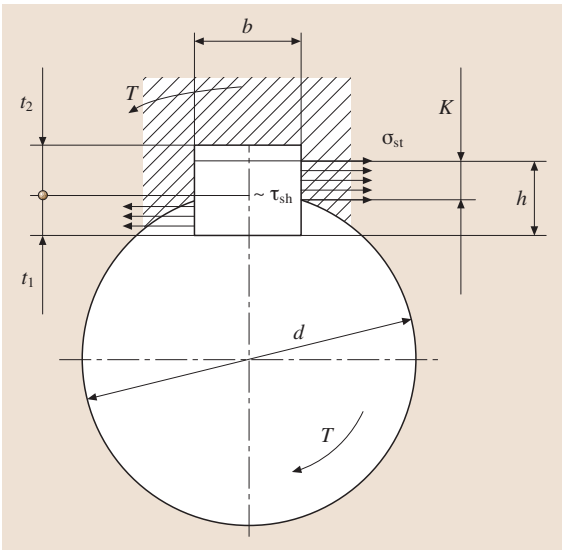


Fig. 6.148 Analytical model of the key joint

The key length $l = l_c + b$ with chamfered $l = l_c$ or with flat ends is chosen from the standard series. The hub length l_{hu} is fixed by ≈ 10 mm more than the key length. To decrease the unevenness of the stress distribution along the height and the length of the key the joint length is limited: $l_{hu} \leq 1.5d$. If the hub length $l_{hu} > 1.5d$ is obtained according to the results of the key joint calculation, so it is advisable to apply a spline connection or an interference connection instead of a key joint.

The strength condition according to shearing stresses is

$$\tau_{sh} = 2 \times 10^3 T/(dbl_c) \leq [\tau]_{sh} ,$$

where b is the key width (mm) and $[\tau]_{sh}$ is the allowable shearing stress (N/mm^2).

A semicircular key represents a disk part with the diameter D and the thickness b. The key height $h \approx 0.4D$ and the length $l \approx 0.95D$.

The groove on the shaft for the semicircular key is made with a disk cutter; in the hub it is made with a broaching cutter or a shaping cutter. Such a fabrication method provides ease of installation and removal of the key, with interchangeability of the connection. A manual fit is usually not needed. The key in the shaft groove self-installs and does not require extra fastening to the shaft.

The disadvantage of such a connection is a weakening of the shaft cross-section with a deep groove, which decreases the fatigue strength of the shaft. Thus, semicircular keys are applied by transmission of relatively low torques. Semicircular keys, like straight ones, work with the *side faces* (Fig. 6.148). The keys are standardized; for every shaft diameter d the values b, h, t_1, t_2 and D are given in the standard. The keys are checked for strength according to bearing stresses σ_{st} and shearing stresses τ_{sh} in compliance with the formulas given for straight keys; $l_c \approx l$.

Materials of Keys and the Choice of Allowable Stresses

Medium-carbon steels with tensile stress $\sigma_t \geq 590$ N/mm^2 serve as the material for the keys (e.g. steel grades E 355 (EN), C 46, C 45 (EN), C 50 E (EN), Appendix 6.A Table 6.95). The values of the allowable stresses of the steel shaft for the key joints are chosen depending on the load condition and operating conditions of the connection from Table 6.66 (the shaft is made of steel).

Higher values are taken under constant load, lower ones are under varying load and operation with impacts. In the case of reverse load $[\sigma]_{st}$ is reduced 1.5 times. The allowable stress on the key shearing is $[\tau]_{sh} = 70–100$ N/mm^2. The higher value is taken under constant load.

Table 6.66 Allowable stresses $[\sigma]_{st}$ for key joints (steel shaft)

Connection type, hub material	$[\sigma]_{st}$ (N/mm^2)
Fixed, steel hub	130–200
Fixed, hub is iron or steel casting	80–110
Sliding without load, steel hub	20–40

The key joint is labor-intensive in manufacture. By the torque transmission considerable local deformations of the shaft and the hub characterize it, which results in uneven pressure distribution on the contact area of the mounting surfaces of the shaft and the hub, as well as on the active faces of the key and the key grooves, which in turn decreases the fatigue shaft strength. Thus, application of the key joints must be limited. They should be used only an interference fit, for the given torque cannot be made in consequence of insufficient material strength of the wheel.

With torque transmission through the key joint, application of the wheel fits on the shaft with clearance is prohibited, and the transition fits are undesirable. If there is a clearance in the connection, the shaft rotation runs with surface slipping of the shaft and the wheel opening, which results in wear-out. This is why the interference should by made with the torque transmission with the key on the mounting surfaces of the shaft and the wheel opening, which guarantees nonopening of the junction.

With the torque transmission with the key joint the fits for the wheels can be assumed according to the following recommendations:

Cylindrical straight	$H7/p6 (H7/r6)$,
Cylindrical helical and worm	$H7/r6 (H7/s6)$,
Bevel	$H7/s6 (H7/t6)$,
Gearboxes	$H7/k6 (H7/m6)$.

The fits with a great interference are given in brackets for the wheels of reverse gears.

For the cases that do not have jointing planes along the shaft axes (e.g. in gearboxes), the choice of the wheel fits is determined by the assembly technique. Assembly is carried out inside the case in the straightened conditions, which is why transition fits are applied for the wheels of gearboxes.

When mounting the gear wheels on the shafts with interference it can be difficult to match the key groove

of the wheel with the shaft key. For ease of installation it is recommended that a guiding cylindrical shaft part with the tolerance range $d11$ (Fig. 6.149a) be foreseen. Sometimes instead of the direction along the cylindrical surface the shaft end is directed to the cone.

Toward the same end the key is beyond the bounds of the component (Fig. 6.149b). With such an execution the length of the shaft slot remains the same. Therefore, the variant in Fig. 6.149b is preferable, although it is more difficult to produce, as the groove for the extension of key is made on the mating component. In either of the two considered variants at first the key groove of the wheel is matched with the key by means of free turning of the wheel relative to the shaft, and then the wheel is pressed onto the shaft.

Mounting surfaces for gear and worm wheels are mostly ground. This is why it is desirable to make a groove for the grinding wheel outlet on the shaft in front of the thrust collars. The key fits are regulated for straight and semicircular keys. The width of the straight key and the thickness of the semicircular one are within the tolerance range $h9$. The following dimensional tolerance ranges are recommended:

- Width of the key groove of the shaft for the straight key: $P9$
- Width of the key groove of the shaft for the semicircular key: $N9$
- Width of the key groove of the hole:
  - With a fixed joint of the irreversible gear: $JS9$
  - With a fixed joint of the reversible gear: $P9$
  - With the sliding joint for the straight key: $D10$

6.10.2 Spline Connections

The cusps (teeth) on the shaft form a spline connection, which enter the corresponding valleys (splines) in the hub. The effective areas are the flanks of the cusps. The cusps on the shaft are made by milling, planning, or rolling in the cold condition with profile rollers according to the method of longitudinal knurling. The valleys in the axle seat are manufactured by broaching or slotting. The spline connections are applied for the fixed joint with the shaft, for the sliding joint along the shaft without load, and the sliding joint under load.

Connections with a straight-sided profile are the most commonly used. They have a constant cusp thickness (Fig. 6.150). The standard foresees three series of connections with a straight-sided profile: *easy*, *medium*, or *difficult*, which differ in height and cusp number z. The difficult series has more elevated cusps with a higher number and is recommended for the transmission of high torques.

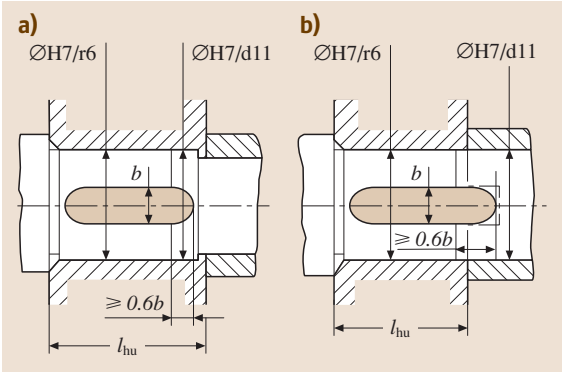


Fig. 6.149a,b Variants with easy assembly of the key joint: **(a)** with a guiding cylindrical section, and **(b)** with a groove in the mating component

Centering (providing coincidence of the geometrical axes) of the connectable components is carried out along the outer D diameter, the inner d diameter, or the lateral faces b. The choice of centering method depends on the requirements for centering accuracy, and the hardness of the hub and the shaft. The two first methods guarantee the most precise centering. The clearance in the contact of the surfaces of the centering method is practically absent, but is considerable in the noncentering one.

Centering Along the Outer Diameter D (Fig. 6.150a)

In this case, machining accuracy of the mating surfaces is provided in the hole through broaching work and on the shaft through grinding. Along the diameter D trim-

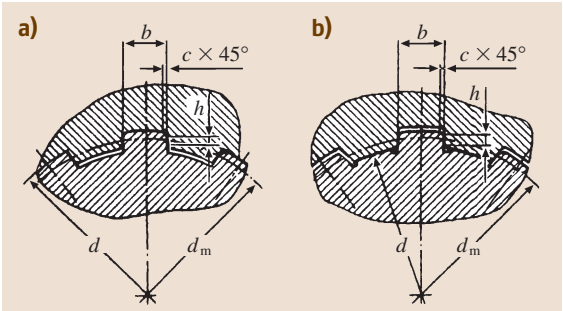


Fig. 6.150a,b Spline connection with a straight-sided profile and centering according to **(a)** the outer diameter or **(b)** the inner diameter

ming is guaranteed on one of the transition fits. Along the inner diameter d there is clearance between the components. By transmission of the torque to the effective lateral faces bearing stresses σ_{st} act. The height of the contact area is

$$h = 0.5(D-d) - 2c\,,$$

where c is the dimension of the bevel.

It is supposed that the peripheral force is applied on the mean diameter of the cusp

$$d_m = 0.5(D+d)\,.$$

In accordance with machining technology for the centering surface in the hole (broaching work), centering along the outer diameter can be used in the case of a low-hardness hub (≤ 350 HB).

Centering along the Inner Diameter d (Fig. 6.150b)

Centering along the inner diameter d is applied with a high-hardness hub (≥ 45 HRC), e.g., after quenching, when calibration of the hub by means of broaching or drift is difficult. The machining accuracy of the mating surfaces is guaranteed in the hole through grinding on the internal grinding machine and on the shaft through grinding of the valley, in accordance with the fact the grooves for the grinding wheel outlet are foreseen.

Along the diameter d, transition fit mating is provided. The dimension h of the contact area is determined in the same way as for outer-diameter centering.

Centering along D or d is used in connections that require high coaxiality of the shaft and the hub (for installation of gear or worm wheels on the shafts in gearboxes of cars, machines, reduction gears, as well as pulleys, chainwheels, and half-couplings on the outlets and the inlets of shafts).

Centering along the Lateral Faces b

In this case, the evident clearance is found in the component mating along the diameters D and d, and the clearance is practically missing on the lateral faces. This decreases the centering accuracy, but guarantees the most even load distribution between the cusps. Thus, centering along the lateral faces b is applied for transmission of considerable torques or those that vary in value or direction, in the case of rigid requirements to the backlash and in the absence of high demands to the centering accuracy, e.g., spline connection of car drive shafts.

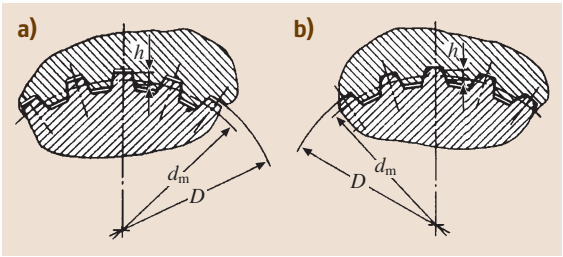


Fig. 6.151a,b Spline connection with involute profile and centering according to **(a)** side faces of the teeth or **(b)** the outer diameter

Connection with an Involute profile (Fig. 6.151)

This approach is applied in fixed and sliding joints. The lateral cusp face is outlined along an involute profile (such as the tooth profile of the gear wheels). The involute profile is distinguished from a straight-sided one by its increased hardness due to cusp thickening and a graded junction at the base. Through cusp manufacture a very good debugged fabrication technique of gear wheel teeth is applied. The joints guarantee a high centering accuracy. The outer diameter D is assumed to be the nominal diameter of the joint. The lower cusp height ($h \approx 1.1m$, where m is the module) and a greater profile angle (here 30°) distinguish them from the teeth of gear wheels, due to a lack of rolling.

In comparison with straight-sided joints, connections with an involute profile are characterized by higher load-carrying capacity as a consequence of their greater contact area, higher tooth number, and increased hardness. They are applied for the transmission of high torques and are considered as promising ones.

Centering is usually done along the tooth surfaces s (Fig. 6.151a), and rarely along the outer diameter D (Fig. 6.151b). Just as in gear wheels the joint parameters are recorded by means of the module m. The mean diameter is $d_m = D - 1.1m$. The height of the contact area for centering along s is $h = 0.9m$ and for centering along D it is $h = m$ ($m = 0.5$–10 mm).

Calculation of Spline Connections

Spline shafts and hubs are manufactured from medium-carbon and alloy steels with tensile stress $\sigma_t > 500\,\mathrm{N/mm^2}$. Failure of the spline connections is caused by damage to the effective surfaces, e.g., wear, bearing, and jamming.

The main criteria for the efficiency of spline connections are bearing and wear resistance of the work surfaces. Wear of the tooth surfaces is caused by microdisplacement of the joint components as a conse-

Table 6.67 Allowable stresses $[\sigma]_{st}$ for spline connections

Type of joint	$[\sigma]_{st}$ (N/mm^2) with hardness	
	≤ 350 HB	≥ 40 HRC
Fixed	60–100	100–140
Sliding without load (pinion unit of a gearbox)	20–30	30–60
Sliding under load (joint of a driveshaft)	–	5–15

quence of resiliency under the action of the radial force and torque, or a mismatch of the rotation axes (because of the presence of clearances, and errors of production and assembly).

Joint parameters are chosen from standard tables depending on the shaft diameter, and then the efficiency criterion calculation is carried out.

To provide the required efficiency a checking analysis is carried out. Bearing and wear of the work surfaces are due to the bearing stresses σ_{st} acting on the surfaces.

The short-cut (approximate) calculation is based on the limitation of the bearing stresses σ_{st} by the allowable values $[\sigma]_{st}$, which are fixed on the basis of field experience with similar structures

$$\sigma_{st} = 2\times 10^3 T K_t/(d_m z h l_c) \le [\sigma]_{st} ,$$

where T is a rated torque (the highest of the long-acting torques under varying loading conditions) (N m), K_t is an irregularity ratio of load distribution between the teeth (that depends on the manufacturing accuracy, the errors of the pitch angles of the cusps and mating sockets, the value of the radial clearance, and the working conditions), $K_t = 1.1–1.5$, d_m is the mean diameter of the joint (mm), z is a cusp number, h is a working cusp height (mm), l_c is a working joint length (mm), and $[\sigma]_{st}$ is the allowable bearing stress (N/mm^2).

In Table 6.67 the values of $[\sigma]_{st}$ are given for products of general engineering and hoist transport systems that are intended for a long lifetime. Higher values are assumed for easy loading conditions.

Through the projection calculation of the spline connections the length of the cusps l_c is determined after the choice of the section dimensions according to the standard. If $l_c > 1.5d$ is obtained, the dimensions and heat treatment are changed or another kind of joint is assumed. The length of the hub is assumed to be $l_{hu} = l_c + (4–6)$mm or more, depending on the joint structure.

Adjusted bearing and wearing calculations are worked out for straight-sided spline connections and load conditions; the design philosophy of the joint, the working surface grind, the required lifetime, etc., are taken into account.

The component fits of the spline connections are regulated by standards. Mostly the fits of the straight-sided splines according to Table 6.68 and involute ones according to Table 6.69 are used.

6.10.3 Pressure Coupling

Load-Carrying Capacity of Pressure Coupling

Interference connections are widely used in practice for the transmission of torque, axial forces, and bending moments. Connections along the cylindrical surfaces are primarily spread. The nature of the joint is such that the shaft is connected to a bushing that has a hole diameter slightly smaller than that of the shaft. At the junction site the components strain elastically and a contact pressure p arises on the surface of the contact, leading to frictional forces on the joint surface that are

Table 6.68 Fits of the elements of straight-sided spline connections

Centering along the surface	Joint	Gear	Surface fits	
			Centering	Lateral
D	Fixed	Irreversible Reversible	$H7/js6$ $H7/n6$	$F8/f7$ $F8/js7$
	Sliding	Irreversible Reversible	$H7/f7$ $H7/h7$	$D9/d9$ $F8/f8$
d	Fixed	Irreversible Reversible	$H7/h7$ $H7/js6$	$H9/h10$ $F10/js7$
	Sliding	Irreversible Reversible	$H7/f7$ $H7/g7$	$H9/d10$ $H9/f9$

Table 6.69 Fits of the elements of involute spline connections

Centering along the surface	Joint	Gear	Surface fits	
			Centering	Noncentering
D	Fixed	Irreversible Reversible	$H7/js6$ $H7/n6$	$9H/9h$
	Sliding	Irreversible Reversible	$H7/g6$ $H7/h6$	$9H/9g$
d	Fixed	Irreversible Reversible	$7H/7n$ $7H/9r$	–
	Sliding	Irreversible Reversible	$9H/8f$ $9H/9g$	–

able to take the external axial force F_a and the torque T (Fig. 6.152). The short-cut calculation for interference joints is based on the assumption that the contact pressures p are evenly distributed along the contact surface.

Pressure couplings are applied to the junctions of gear and worm wheels, pulleys, chainwheels, inner races of rolling bearings, rotors of electric motors with a shaft, and to junctions of rings of gear and worm wheels, etc., and disks. They are used in the production of composite crankshafts, units of drive chains, junctions of railway wheels with axes, and tires. Component connections with interference are considered permanent connections, as they allow a limited amount of dismantling and reassembly.

Cylindrical joints, depending on the installation method, are divided into those that are mounted by means of insertion and heating of the female part (bushing) or those that are mounted by means of cooling of the male part (shaft).

Component insertion is carried out on hydraulic, screw, or lever presses. To prevent tearing and to decrease the insertion forces the mating surfaces are lubricated with oil.

The reheating temperature must be lower than that of low-temperature tempering, so that structural changes do not occur in the metal. For cooling of the shaft carbonic acid or liquid air is used.

At present, so-called *thermomechanical* component *joints* are used more often, being produced from *shape-memory* alloys (iron–manganese, copper–aluminum–nickel, copper–zinc–aluminum, etc.), which are characterized by the possibility of diffusion-free transformation of one solid solution into another as a result of mechanical effects or thermal action. This property is particular to, e.g., nickel–titanium alloys, which experience *reversible martensitic transformation*. It characterizes the ability of the material, which is deformed in the martensitic state (by low temperature), to recover partially ($\approx 25\%$) obtained inelastic deformation during the following heating and transition into the austenitic state.

To form the thermomechanical joint, e.g., the bushing is manufactured from a shape-memory alloys. It is cooled in liquid nitrogen (so the material is in a martensitic state), the hole is strained with a drift in the radial direction to the formation of technological clearance by the following installation of the bushing on the mating shaft. Assembly is carried out at ambient temperature. Subsequent heating of the bushing with ambient heat results in the corresponding recovery of the previous hole dimensions as a consequence of the transition of the material into the austenitic state, leading to the formation of interference in the joint. Such joints are widely applied in aeronautical and space engineering, nuclear-power engineering, and medicine (automatic relays, space antenna drive mechanisms, and self-sealed joints).

The male part (shaft) is given the index 1, and the female part (bushing) has the index 2. A bushing is interpreted as any component installed on the shaft: a hub

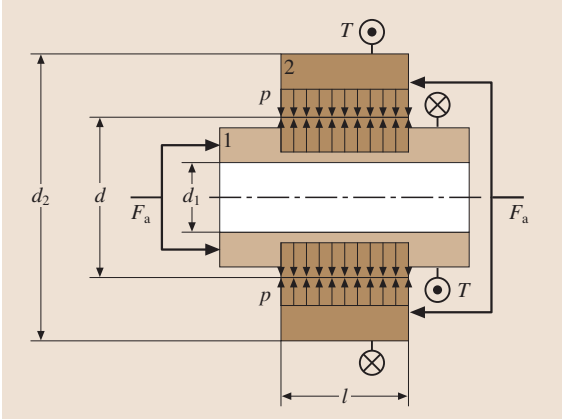


Fig. 6.152 Diagram for the determination of the load-carrying capability of interference joints

of gear or worm wheels, pulleys, chainwheels, and the inner race of bearings, etc.

The efficiency conditions of the interference connection are the lack of relative displacement of the components under the action of the axial force F_a and the lack of relative component turning under the action of the torque T.

The shafts rotate relative to the loads acting on them. This is why stresses change cyclically in some ranges at any contact point per revolution of the shaft. Stress cycling results in the effect of surface layer fatigue of the component material, in microslip of the mounting surfaces, and as a result, in wear, i. e., in so-called contact corrosion. The interference of the joint in this case is progressively decreased and there comes a point at which the hub turns relative to the shaft.

To prevent contact corrosion, or to reduce its influence in interference connections, a definite traction reserve K should be included, which is assumed to be:

- For the wheels of the output shafts of reduction gears, on whose ends there is

  A joint sleeve: $K = 3$
  A chain wheel: $K = 3.5$
  A belt drive sheave: $K = 4$

- For the wheels of the countershafts of reduction gears: $K = 4.5$

Calculation of Pressure Coupling Loaded with a Torque and an Axial Force

The purpose of the calculation is to match the interference fit. The basic data are as follows: required for transmission rotatory torque T (N m), axial force F_a (N), as well as d – the joint diameter (mm), d_1 – the diameter of the central axial hole of the shaft (mm), d_2 – the passage outer diameter of the bushing (of the wheel hub, the outer diameter of the rim, etc.) (mm), l – the mating length (mm), and materials of the connecting components, and the roughness parameters of the mating surfaces.

Matching of the fit is carried out in the following order:

1. *Average contact pressure* (N/mm^2)

$$p = KF/(\pi dlf) ,$$

where K is a traction safety factor; $F = \sqrt{F_a^2 + F_t^2}$ is a relative total peripheral force (N), $F_t = 2\times10^3 T/d$ is a peripheral force (N), and f is a traction coefficient (friction).

Table 6.70 Recommended values of the traction coefficient f

Material pair	f when mounting by	
	Insertion	Heating
Steel–iron	0.08	0.14
Steel–steel	0.08	0.14
Steel–bronze (brass)	0.05	0.07
Iron–bronze (brass)	0.05	0.07

2. *Component deformation* (μm)

$$\delta = 10^3 pd(C_1/E_1 + C_2/E_2) ,$$

where C_1, C_2 are stiffness factors

$$C_1 = [1 + (d_1/d)^2]/[1 - (d_1/d)^2] - \nu_1 ,$$
$$C_2 = [1 + (d/d_2)^2]/[1 - (d/d_2)^2] + \nu_2 ,$$

where E is a coefficient of elasticity (N/mm^2), being for steel 2.1×10^5, iron 0.9×10^5, tin bronze 0.8×10^5, for tinless bronze, and for brass 10^5; ν is Poisson's ratio: being for steel 0.3, for iron 0.25, and for bronze and brass 0.35.

3. *Allowance for pressing down of microasperities* (μm)

$$u = 5.5(Ra_1 + Ra_2) ,$$

where Ra_1 and Ra_2 are the arithmetic-mean deviations of the surface profile of the shaft and the hole, respectively. Generally, $Ra_1 = 0.8$ μm and $Ra_2 = 1.6$ μm.

4. *Allowance for thermal deformation* (μm)
By fit matching in mating of the components, which heat in operation to relatively high temperatures, thermal deformations that loosen the interference are calculated according to the formula

$$\delta_t = 10^3 d[(t_2 - 20\,^\circ\mathrm{C})\alpha_2 - (t_1 - 20\,^\circ\mathrm{C})\alpha_1] .$$

Here t_1 and t_2 are the average volumetric temperatures of the shaft and the bushing, respectively; α_1, α_2 are the linear expansion coefficients (1/°C) of the shaft and the bushing, respectively, being for steel $\alpha = 12\times10^{-6}$, for iron $\alpha = 10\times10^{-6}$, and for bronze and brass $\alpha = 19\times10^{-6}$.

5. *Minimum interference* (μm), required for load transmission,

$$[N]_{\min} = \delta + u + \delta_t .$$

6. *Maximum interference* (μm), permissible with the component strength (of the hub, ring, etc.),

$$[N]_{\max} = [\delta]_{\max} + u .$$

Here $[\delta]_{max} = [p]_{max}\delta/p$ (μm) is the *maximum deformation* permissible for the strength of the joint components, where $[p]_{max}$ (N/mm^2) is the maximum pressure allowable for the strength of the male or female part, being the lowest of

$$[p]_{max2} = 0.5\sigma_{y2}\left[1-(d/d_2)^2\right] \quad \text{or}$$
$$[p]_{max1} = 0.5\sigma_{y1}\left[1-(d_1/d)^2\right] .$$

For a shaft without a central hole ($d_1 = 0$), one has $[p]_{max\,1} = \sigma_{y1}$, where σ_{y1} and σ_{y2} are the yield strengths of the material of the male and female parts (N/mm^2), respectively.

7. *Fit choice*
According to the values $[N]_{min}$ and $[N]_{max}$ the fit is chosen from Table 6.73, to meet the conditions $N_{min} \geq [N]_{min}$ and $N_{max} \leq [N]_{max}$.
In Table 6.73 the values of the minimum N_{min} and maximum N_{max} stochastic interferences (for the usually assumed probability of 99.73%) are determined from the formulas, taking into account the dimensional dispersion of the shaft and the hole and as a result of interference dispersion.
8. For the chosen fit, the insertion force or the reheating temperature of the component are determined
Insertion force (N)

$$F_s = \pi dl\, p_{max} f_s ,$$

where $p_{max} = (N_{max} - u)p/\delta$ (N/mm^2) is the pressure from interference N_{max} of the chosen fit and f_s is a traction coefficient (friction) due to insertion:

Table 6.71 Recommended values of traction coefficient f_S due to insertion

Material pair	f_s
Steel–steel	0.20
Steel–iron	0.14
Steel–bronze (brass)	0.10
Iron–bronze (brass)	0.08

Reheat temperature of the female part °C

$$t = 20°\text{C} + (N_{max} + Z_a)/(10^3 d\alpha_2) ,$$

where Z_a is a clearance (mm); for convenience of assembly it is assumed to depend on the shaft diameter d:

Table 6.72 Recommended clearance Z_a values

d (mm)	Z_a (μm)
Over 30 to 80	10
Over 80 to 180	15
Over 180 to 400	20

The reheating temperature must be such that there are no structural changes in material. The allowable temperature for steel is $[t] = 230–240$ °C while for bronze it is $[t] = 150–200$ °C.

Interference Connections with Electroplated Coatings

The load-carrying capacity of interference connections can be considerably increased by covering the electroplated coating on the mounting surfaces. Figure 6.153 shows the results of comparison testing of interference joints. Electroplated coatings were layered on the mounting surfaces with a thickness ranging from 0.01 to 0.02 mm. The joints were assembled by means of two methods: insertion (columns a) and with shaft cooling in liquid nitrogen (columns b). In the latter case clearance of 0.05 mm to the side formed between connected surfaces when assembled. The displacement force F_0 is assumed as a unit of comparison for the control junction without coating, which is assembled by means of insertion (without shaft cooling). It is evident

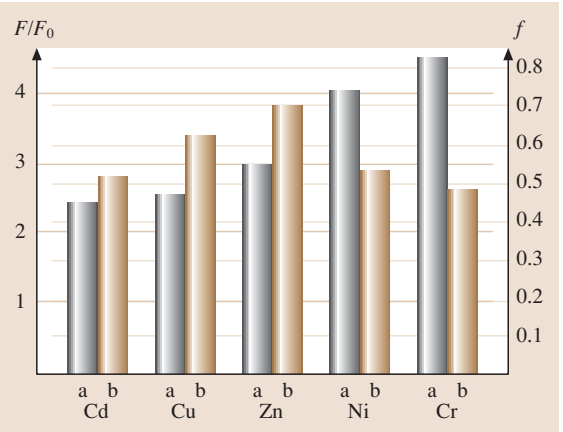


Fig. 6.153 Load-carrying capability of interference joints with electroplated coats. a – Assembly by means of insertion. b – Assembly with cooling of the shaft in liquid nitrogen

Table 6.73 Stochastic minimum N_{min} and maximum N_{max} interferences of fits

Diameter d (mm)	Interference values $\frac{N_{min}}{N_{max}}$ (μm) for the fits										
	$\frac{H7}{p6}$	$\frac{H7}{r6}$	$\frac{H8}{s7}$	$\frac{H7}{s6}$	$\frac{H7}{s7}$	$\frac{H7}{t6}$	$\frac{H8}{u8}$	$\frac{H7}{u7}$	$\frac{H8}{x8}$	$\frac{H8}{z8}$	$\frac{H8}{za8}$
Over 30 up to 40	$\frac{7}{36}$	$\frac{15}{44}$	$\frac{13}{59}$	$\frac{24}{53}$	$\frac{25}{61}$	$\frac{29}{58}$	$\frac{32}{88}$	$\frac{42}{78}$	$\frac{52}{108}$	$\frac{84}{140}$	$\frac{120}{175}$
Over 40 up to 50	$\frac{7}{36}$	$\frac{15}{44}$	$\frac{13}{59}$	$\frac{24}{53}$	$\frac{25}{61}$	$\frac{35}{64}$	$\frac{42}{98}$	$\frac{52}{88}$	$\frac{69}{125}$	$\frac{108}{164}$	$\frac{152}{207}$
Over 50 up to 65	$\frac{9}{44}$	$\frac{18}{53}$	$\frac{18}{72}$	$\frac{30}{65}$	$\frac{32}{74}$	$\frac{43}{78}$	$\frac{55}{119}$	$\frac{66}{108}$	$\frac{90}{154}$	$\frac{140}{204}$	$\frac{193}{258}$
Over 65 up to 80	$\frac{9}{44}$	$\frac{20}{55}$	$\frac{24}{78}$	$\frac{36}{71}$	$\frac{38}{80}$	$\frac{52}{87}$	$\frac{70}{134}$	$\frac{81}{123}$	$\frac{114}{178}$	$\frac{178}{242}$	$\frac{241}{306}$
Over 80 up to 100	$\frac{10}{51}$	$\frac{24}{65}$	$\frac{29}{93}$	$\frac{44}{85}$	$\frac{46}{96}$	$\frac{64}{105}$	$\frac{86}{162}$	$\frac{99}{149}$	$\frac{140}{216}$	$\frac{220}{296}$	$\frac{297}{373}$
Over 100 up to 120	$\frac{10}{51}$	$\frac{27}{68}$	$\frac{37}{101}$	$\frac{52}{93}$	$\frac{54}{104}$	$\frac{77}{118}$	$\frac{106}{182}$	$\frac{119}{169}$	$\frac{172}{248}$	$\frac{272}{348}$	$\frac{362}{438}$
Over 120 up to 140	$\frac{12}{59}$	$\frac{32}{79}$	$\frac{43}{117}$	$\frac{61}{108}$	$\frac{64}{120}$	$\frac{91}{138}$	$\frac{126}{214}$	$\frac{142}{193}$	$\frac{204}{292}$	$\frac{320}{410}$	$\frac{425}{514}$
Over 140 up to 160	$\frac{12}{59}$	$\frac{34}{81}$	$\frac{51}{125}$	$\frac{69}{116}$	$\frac{72}{128}$	$\frac{103}{150}$	$\frac{155}{243}$	$\frac{171}{227}$	$\frac{236}{324}$	$\frac{370}{460}$	$\frac{490}{579}$
Over 160 up to 180	$\frac{12}{59}$	$\frac{37}{84}$	$\frac{59}{133}$	$\frac{77}{124}$	$\frac{80}{136}$	$\frac{115}{162}$	$\frac{166}{254}$	$\frac{182}{238}$	$\frac{266}{354}$	$\frac{420}{510}$	$\frac{555}{644}$
Over 180 up to 200	$\frac{14}{69}$	$\frac{41}{95}$	$\frac{66}{152}$	$\frac{86}{140}$	$\frac{89}{155}$	$\frac{130}{184}$	$\frac{185}{287}$	$\frac{203}{269}$	$\frac{299}{401}$	$\frac{469}{571}$	$\frac{619}{721}$
Over 200 up to 225	$\frac{14}{69}$	$\frac{44}{98}$	$\frac{74}{160}$	$\frac{94}{148}$	$\frac{97}{163}$	$\frac{144}{198}$	$\frac{207}{309}$	$\frac{225}{291}$	$\frac{334}{436}$	$\frac{524}{626}$	$\frac{689}{791}$
Over 225 up to 250	$\frac{14}{69}$	$\frac{47}{101}$	$\frac{84}{170}$	$\frac{104}{158}$	$\frac{107}{173}$	$\frac{160}{214}$	$\frac{233}{335}$	$\frac{251}{317}$	$\frac{374}{476}$	$\frac{589}{691}$	$\frac{769}{871}$
Over 250 up to 280	$\frac{15}{77}$	$\frac{53}{115}$	$\frac{95}{191}$	$\frac{117}{179}$	$\frac{121}{195}$	$\frac{177}{239}$	$\frac{258}{372}$	$\frac{278}{352}$	$\frac{418}{532}$	$\frac{653}{767}$	$\frac{863}{977}$
Over 280 up to 315	$\frac{15}{77}$	$\frac{57}{119}$	$\frac{107}{203}$	$\frac{129}{191}$	$\frac{133}{207}$	$\frac{199}{261}$	$\frac{293}{407}$	$\frac{313}{387}$	$\frac{468}{582}$	$\frac{733}{847}$	$\frac{943}{1057}$

from Fig. 6.153 that covering increases the displacement force by a factor of 2–4.5. The load-carrying capacity of joints assembled by means of shaft cooling exceeds the strength of the connection assembled by means of insertion in by a factor of 2 for joints without coating and by factors of 1.2–1.3 for joints with soft coatings (Cd, Cu, and Zn). For joints with hard coatings (Ni and Cr) the load-carrying capacity is lower when assembled with cooling than when assembled with insertion.

The traction increase by electroplated coatings is caused by interdiffusion in the case of high pressure of the coating and parent metal, which is accompanied by the formation of the intermediate structures. This explains the close approximation to one value of the traction coefficient f (the right ordinate in Fig. 6.153), which actually represents the shearing resistance of the intermediate metal layer.

Application of soft coatings and assembly with shaft cooling increase the load-carrying capacity of the joints by a factor of 3–4 in comparison with joints without coating assembled by means of insertion. Consequently, for a set external load there is the possibility of using fits with lower interferences and corresponding lower tension stresses in the female part (bushing) and compression in the male part (shaft). Moreover, electroplated coatings protect contact surfaces from corrosion and avoid welding.

Calculation of Interference Connections Loaded with a Bending Moment

In some cases, the interference joints, e.g., the connections of gear wheels with shafts, are subjected to loading by a bending moment.

Considering the shaft to be absolutely hard, it can be imagined that the shaft rotates around the axis, which is

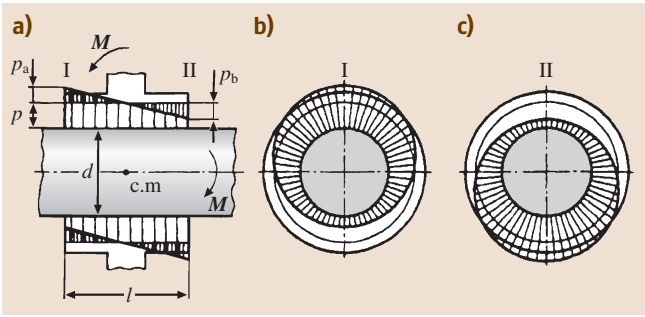


Fig. 6.154a–c Analytical model of connections with interference loaded with a bending moment. **(a)** Summation of diagrams of pressure from interference and bending moment, **(b,c)** the sickle-shaped nature of the total stresses on the faces of the components I and II, respectively

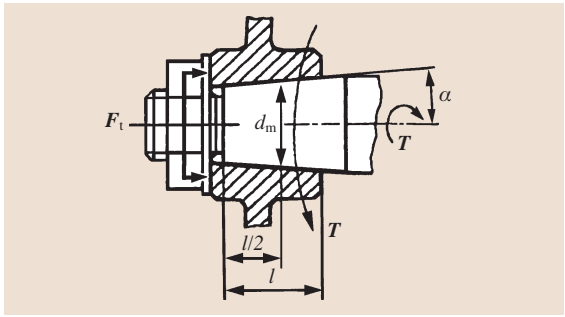


Fig. 6.155 Connection by means of fit on the tapered shaft section

perpendicular to the drawing plane, and passes through the center of mass (c.m.) of the joint (Fig. 6.154a), relative to the bushing under the action of the bending moment M.

By loading of the interference joint with a bending moment M, the pressure p_b diagram, which is characteristic for bending (Fig. 6.154a), is overlaid on the pressure p diagram after the fit. As a result, the pressure distribution changes along the joint length, i. e., there is an increase in the compression zone and a decrease in the tension zone. It is known that the upper part of the joint receives half of the bending moment and the lower part receives the other half. In the cross-sections of the joint, e.g., on ends I and II of the hub, a circular pressure diagram first takes on a sickle-shaped nature (Fig. 6.154b,c). A pressure change does not diminish the ability of the joint to support an axial force and a torque, as the total value of the frictional forces is constant.

The efficiency condition of the joint is the nonopening of the junction in the most unloaded contact zone. Without opening of the junction the residual pressure is assumed to be equal to $p_{\min} = 0.25p$, where p is a pressure from the fit (N/mm^2)

$$p = 10^{-3}\delta/[d(C_1/E_1 + C_2/E_2)] ,$$

where $\delta = N_{\min} - u$.

The highest pressures p_b (N/mm^2) from the bending moment M (N m) are determined by analogy with the bending formula

$$p_b = \frac{10^3 M}{2W}\frac{4}{\pi} \le 0.75p ,$$

where $M/2$ is the part of the bending moment that falls on the upper (or lower) half of the joint, $4/\pi$ is a factor taking into account the sickle-shaped nature of the summary pressure diagram along the circle of the journal, and $W = dl^2/6$ is a bending modulus of the diametral axial section of the journal (mm^3).

After transformation the formula for calculation of the allowable bending moment (N m) is obtained from the prevention condition of the junction opening

$$M \le 0.2\times 10^{-3} dl^2 p .$$

Obviously, the allowable bending moment is proportionate to the squared length, which is why it is advisable to increase the length l when increasing the load-carrying capacity of a joint that is subjected to the action of the considerable bending moment.

Fit Connection on a Conical Shaft Part

Fit connections on a conical shaft part are mainly applied to fasten the components onto the shaft ends (Fig. 6.155). Interference and contact pressures are formed, e.g., with nut tightening that loads the connection with an axial force F_t. In contrast to cylindrical connections, bevel connections are easily installed and dismantled without special tools. Other advantages over cylindrical connections are: precise centering, the possibility of checking the interference through the axial displacement or pull force, the possibility of repeated assembly and dismantling, as well as retorques in the case of interference slackening in operation. These connections are considered promising ones, which widens their application field.

The pressure p (N/mm^2) on the work surface due to the axial pull force F_t (N) is

$$p = F_t/[\pi d_m l(\tan\alpha + f)] ,$$

Table 6.74 Basic characteristics of conic interference-fit rings

d (mm)	D (mm)	L (mm)	l (mm)	F_{t1} (kN)	F_{t2} (kN)	T (N m)	F_a (kN)
12E7	15f7	4.5	3.7	6.95	7.5	10	1.67
14E7	18f7	6.3	5.3	11.20	12.6	19.6	2.80
15E7	19f7	6.3	5.3	10.75	13.5	22.5	3.00
16E7	20f7	6.3	5.3	10.10	14.4	25.5	3.19
18E7	22f7	6.3	5.3	9.10	16.2	32.4	3.60
20E7	25f7	6.3	5.3	12.05	18.0	40	4.00
22E7	26f7	6.3	5.3	9.05	19.8	48	4.40
24E7	28f7	6.3	5.3	8.35	21.6	58	4.80
25E7	30f7	6.3	5.3	9.90	22.5	62	5.00
28E7	32f7	6.3	5.3	7.40	25.2	78	5.60
30E7	35f7	6.3	5.3	8.50	27.0	90	6.0
32E7	36f7	6.3	5.3	7.85	28.8	102	6.4
35E7	40f7	7.0	6.0	10.10	35.6	138	7.9
36E7	42f7	7.0	6.0	11.60	36.6	147	8.2
38E7	44f7	7.0	6.0	11.00	38.7	163	8.6
40E8	45e8	8.0	6.6	13.80	45.0	199	9.95
42E8	48e8	8.0	6.6	15.60	47.0	219	10.4
45E8	52e8	10.0	8.6	28.20	66.0	328	14.6
48E8	55e8	10.0	8.6	24.60	70.0	373	15.6
50E8	57e8	10.0	8.6	23.50	73.0	405	16.2
55E8	62e8	10.0	8.6	21.80	80.0	490	17.8
56E8	64e8	12.0	10.4	29.40	99.0	615	22.0
60E8	68e8	12.0	10.4	27.40	106.0	705	23.5
63E8	71e8	12.0	10.4	26.30	111.0	780	24.8
65E8	73e8	12.0	10.4	25.40	115.0	830	25.6
70E8	79e8	14.0	12.2	31.00	145.0	1120	32.0

where d_m and l are, respectively, the mean diameter and the joint length (mm), f is a traction coefficient (friction) ($f \approx 0.12$), and α is a gradient angle of the cone generatrix to the shaft axis.

For the shaft ends the taper 1 : 10 is the most commonly used, $\alpha = 2°51'45''$, $\tan\alpha = 0.05$. The torque T (N m) which the connection can transmit is

$$T \le 0.5 \times 10^{-3} F_t d_m f_r ,$$

where f_r is a surface friction factor $f_r = f/(\tan\alpha + f)$.

The required pull force for transmission through the joint for a set torque T becomes

$$F_t = 2 \times 10^3 KT/(d_m f_r) ,$$

where $K = 1.3–1.5$ is a traction safety factor.

Along with these tightening joints, tapered connections are used in dead joints and rarely in dismantled structures, where the interference is formed without application of thread pieces, but, e.g. by, means of insertion with a standardized impact or insertion on the rated axial displacement, or by means of heating of the female part (cooling of the male part). Recommended tapers for such connections are 1 : 50 to 1 : 30.

6.10.4 Frictional Connections with Conic Tightening Rings

Frictional connections with conic tightening rings are used for the installation of components such as gear wheels, pulleys, chainwheels, and half-couplings on shafts.

These connections transmit torques and axial forces due to the frictional forces on the contact surfaces of the shaft and the hub with conic rings installed in the annular gap between the shaft and the hub (Fig. 6.156). In Russia the rings are produced from spring steel 55 Si (EN), etc. (Appendix 6.A Table 6.95). By tightening the nut on the shaft (Fig. 6.156a) or the screw in the hub (Fig. 6.156b) the conic rings are elastically deformed and pull against one another. Then the outer rings are

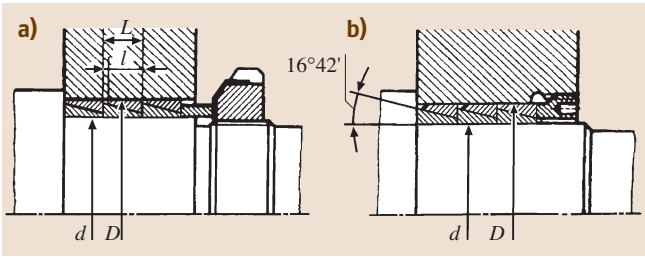


Fig. 6.156a,b Tightening methods of conic rings: **(a)** with a nut on the shaft or **(b)** with a screw in the hub

strained and tightly pressed to the hub, and the inner rings are compressed and tightly pressed to the shaft, leading to interference on the surface of the shaft and the hole.

The connections allow *hub* mounting on the shaft in any angular and axial positions, provide easy assembly, dismantling, good positioning, and tightness, do not loosen the shaft section with grooves or turnings, but need precise ring production with compliance to strict coaxiality of their outer and inner surfaces.

The required tightening force of the ring set for assembling is determined from the formula

$$F_t = F_{t1} + F_{t2} ,$$

where F_{t1} is a force required for the ring deformation when eliminating the mounting clearances; F_{t2} is a force required for formation of the mounting pressure on the shaft, which is equal to $100\,\mathrm{N/mm^2}$.

Table 6.74 shows dimensions of the rings, and values of the axial tightening forces F_{t1} and F_{t2}, transmitted turning moments T, and axial forces F_a for contact pressure of $p = 100\,\mathrm{N/mm^2}$. For $p = 200\,\mathrm{N/mm^2}$ the values of T and F_a are doubled, while for $p = 50\,\mathrm{N/mm^2}$ they are halved. Pressures are chosen depending on the toughness and sliding strength of the contacting surfaces. The values of T and F_a given in Table 6.74 correspond to the traction coefficient (friction) on the mating surfaces $f = 0.12$.

When mounting some ring sets it should be taken into account that a portion of the tightening force overcomes the axial constituents of the friction forces on the ring surfaces. Therefore the moment or the axial force transmitted with the second set amounts to approximately one-half, with the third set it amounts to one-quarter, and with the fourth one it amounts to one-eighth of the nominal values given in Table 6.74.

Tolerance ranges of the mounting surfaces are as follows:

Table 6.75 Recommended tolerance ranges

	Shaft	Hole
$d(D) \le 38\,\mathrm{mm}$	$h6$	$H7$
$d(D) \ge 40\,\mathrm{mm}$	$h8$	$H8$

Example

Select the parameters of the connection with the conic rings for transmission of the torque $T = 700\,\mathrm{N\,m}$ from the spur gear for a shaft with a diameter of 50 mm.

Solution

It follows from Table 6.74 that one ring set with diameter $d = 50\,\mathrm{mm}$ can transmit a torque of $405\,\mathrm{N\,m}$. The second set installed next to the first one transmits half of the load, $202.5\,\mathrm{N\,m}$; the third one transmits $101.25\,\mathrm{N\,m}$. Thus, three ring sets provide the torque transmission: $405 + 202.5 + 101.25 = 708.75\,\mathrm{N\,m}$, which guarantees transmission of the set torque in the situation.

The tightening force depends on the set number and for a diameter of 50 mm (Table 6.74) it is

$$F_t = F_{t1} + F_{t2} = 23.5 + 73 = 96.5\,\mathrm{kN} .$$

6.11 Rolling Bearings

6.11.1 Introduction

A bearing is a support or a guide determining a position of movable parts relative to other mechanism components. Bearings working mainly for movement with rolling friction are called *rolling bearings*, and those for movement with sliding friction are called *friction bearings*. Rolling bearings include details with rolling paths and solids of revolution [6.72, 99–105].

The advantages of rolling bearings are the following:

1. Complete interchangeability, readiness to operate without additional adjustment or debugging.
2. Small axial dimensions, simplicity of assembly and operation.
3. Low need for lubrication. Bearings with safety washers or integral seal are filled with a semisolid lubricant at production. This reserve is sufficient for their entire lifetime.

4. Low losses of friction, especially for breakaway and low rotational frequencies, slight heating under operation.
5. Low use of scarce nonferrous metals during manufacture.
6. Low production costs due to mass manufacturing.

The disadvantages of rolling bearings are the following:

1. Large radial dimensions
2. Axial and radial rigidity are low and vary according to the rotary angle
3. High rotation resistance, noise, and short lifetime at high rotational frequencies
4. Sensitivity to impact and vibrational loads

Application

Rolling bearings are the main type of supports in machines: in cars there are more than 30 types of bearings, in trucks there are more than 120, and in the airplanes, more than 1000, etc.

6.11.2 Classifications of Rolling Bearings

Rolling bearings transmit forces between the shaft and the case by their relative rotation. The forces loading the bearing are divided into:

- *Radial* forces acting in the direction perpendicular to the bearing axis
- *Axial* forces acting in the direction parallel to the bearing axis

Rolling bearings are classified according to the following main features:

- According to the form of the solids of revolution (Fig. 6.157) ball (a) and roller (b–h), where the latter can be designed with rollers that are cylindrical short (b), long (c), needle (d), taper (e), and cambered (f) with a small (7–30 μm side) convexity of the rolling surface (camber) and spiral hollow (g).
- According to the direction of the supported load: radial, to be loaded by radial forces; some types can also receive axial forces; radial-thrust, to be loaded by radial and axial forces; adjustable-type bearings cannot work without axial force; thrust, to be loaded by axial forces, they do not take radial force; thrust-radial, to be loaded by axial and low radial forces.
- According to the row number of the solids of revolution: single row, double row, and four row.
- According to the main design features: self-installed (e.g., spherical bearings self-install by angular displacement of the shaft axes and the hole in the case) and non-self-installed, with cylindrical or conic holes, binary, etc.

Division of bearings depending on the action direction of the supported load is relative in a number of cases. For example, a widespread ball radial single-row bearing is successfully applied for perception not only combined (radial and axial loads acting together), but also merely axial loads, and thrust-radial bearings are usually used only for bearing axial loads. In addition to these main types of bearings, various modifications are also produced.

6.11.3 Main Types of Bearings

The ball radial single-row bearing (Fig. 6.158a) is intended for load transfer of radial and limited axial forces of any direction, is one of the most common and cheap types of bearings. Its load rating is lower than that of roller bearings of the same dimensions. It can operate under the action of axial force only at high rotational frequency, i. e., in conditions for which flat-thrust bearings are not suitable.

It guarantees axial fixing of the shaft in two directions. Being non-self-installed it allows small angles of mutual warp of the inner and outer races, the values of which depend on the radial clearances in the bearing. With the same overall dimensions these bearings oper-

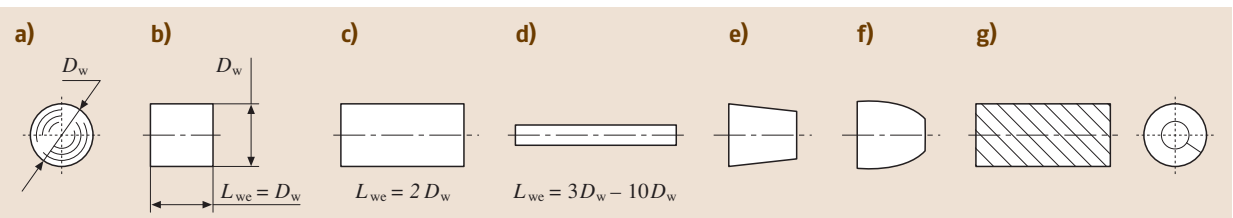


Fig. 6.157a–g Principal types of solids of revolution. **(a)** Ball, **(b)**–**(g)** rollers: **(b)** short cylindrical, **(c)** long cylindrical **(d)** and needle, **(e)** tapered, **(f)** cambered, **(g)** spiral

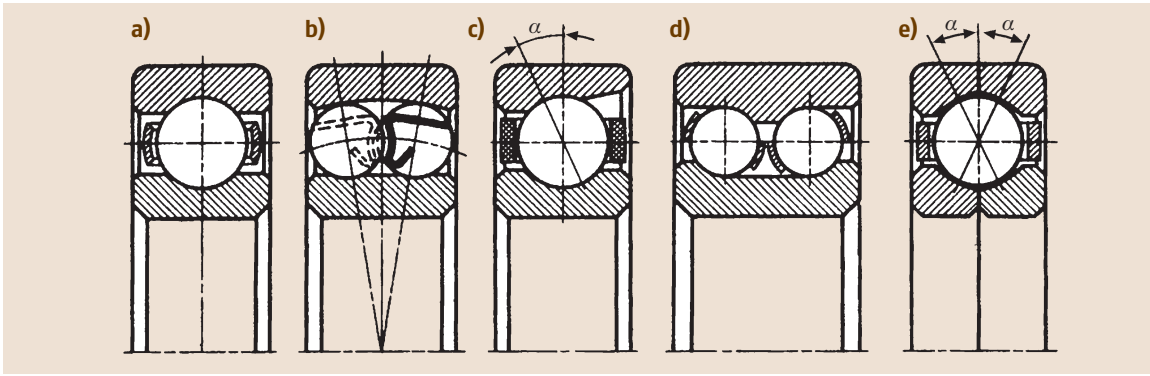


Fig. 6.158a–e Ball bearings. **(a)** Radial single-row, **(b)** radial double-row spherical, **(c)** radial-thrust, **(d)** radial-thrust double-row, **(e)** radial-thrust with a split inner race

ate with lower friction losses and with higher rotational shaft frequency than bearings of other structures.

The ball radial spherical double-row bearing (Fig. 6.158b) is intended for perception of radial forces, but it can also receive limited axial forces in any direction. The presence of the axial constituent results in an imbalance of the force distribution between the rows. The radial dynamic load rating is lower than that of radial single-row bearings.

The rolling path on the outer race is machined along the circle. This is why the bearing is able to self-install and to operate at considerable (up to 1.5−4°) warp of the inner race (of the shaft) relative to the outer race (of the case). It is used in units with nonrigid shafts and in structures in which an appropriate coaxiality of the holes in the cases cannot be provided.

A ball radial-thrust single-row bearing is shown in Fig. 6.158c. The main embodiments differ in the initial contact angles: $\alpha = 12$, 15, 26, and 36°. The higher α, the higher the axial rigidity and the dynamic load rating, but the lower the allowable radial load and limit rotation frequency (because of the influence of gyroscopic effects). This bearing is intended for bearing radial and axial forces of one direction only; it cannot run under the action of radial force only without axial force.

For the perception of axial forces of any direction and for double-sided shaft fixing these bearings are installed on the shaft in pairs. On assembly, the unit the bearings should be adjusted for approximately zero clearance between the balls and the grooves of the races under steady thermal conditions. In some machines (e.g., machine tools), at adjustment, pair bearings are assembled with prior interference, whereby the bearing rigidity and rotational accuracy increase.

Radial-thrust bearings differ from radial bearings in that they have a greater number of balls that can be positioned in the bearing as a consequence of the presence of the chamfer on the outer race. Thus, their rigidity and dynamic load rating are higher.

The ball radial-thrust double-row bearing (Fig. 6.158d) is intended for bearing considerable radial, axial, and combined loads with high rigidity requirements. They are manufactured with a prior interference.

The ball radial-thrust single-row bearing with a split-face inner (or outer) race and a contact at three or four points (Fig. 6.158e) is intended for bearing radial and double-sided loads in the conditions of restricted dimensions in the axial direction.

The main embodiment of the ball thrust bearing is a single bearing (Fig. 6.159a). It is intended for bearing axial force of only one direction. It runs worse on horizontal shafts than on vertical ones and requires good adjustment or constant tightening of the races with the springs. One bearing race is installed on the shaft according to the interference fit. It is very sensitive to assembling accuracy. Because of increased gyroscopic effects it is used at considerably lower rotational frequencies than other ball bearings.

Other embodiments include:

- Single with a free race self-installed along the spherical outer surface and a jar washer (Fig. 6.159b); the jar washer compensates for any lack of parallelism between the supporting surfaces of the case and the collar of the shaft.
- Double thrust bearing with three races (Fig. 6.159c); the central race fixed onto the shaft has two grooves, it is used for bearing axial forces in both directions.

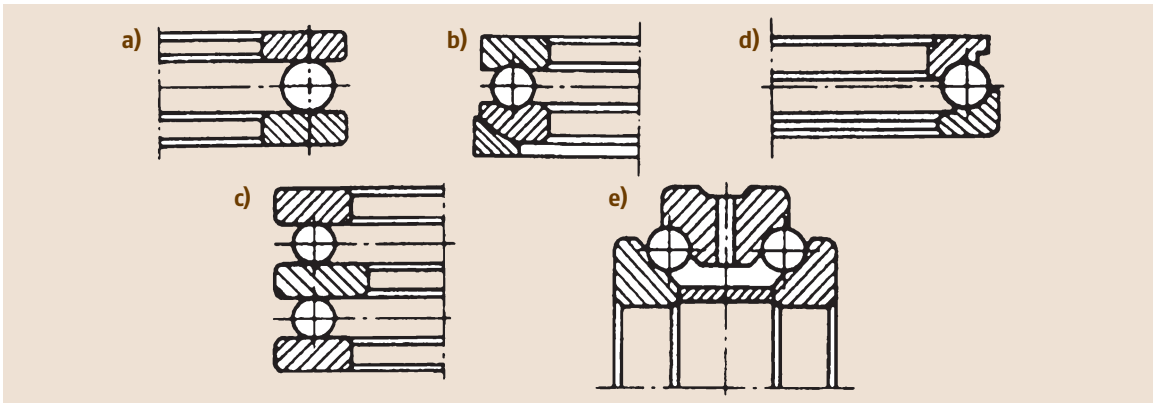


Fig. 6.159a–e Ball bearings. **(a)** Thrust single, **(b)** thrust with a spherical jar washer, **(c)** thrust double, **(d)** thrust-radial single, **(e)** thrust-radial double

The ball thrust-radial bearing allows higher rotational frequencies than does the thrust bearing. It is less sensitive to mutual warp of the races than the ball thrust bearing.

Embodiments of the ball thrust-radial bearings include:

- Single (Fig. 6.159d), used for bearing radial and axial forces in one direction.
- Paired with a contact angle of 60° (Fig. 6.159e), for which the direction of the forces is axial on both sides and radial.

The roller radial single-row bearing with short cylindrical rollers of the main embodiment, with ledges on the inner race and without ledges on the outer race (Fig. 6.160a), can bear only radial force. Roller bearings are noted for their higher dynamic load rating than that of ball bearings. Split assembly of the inner (with a roller set) and outer races is possible, which allows axial mutual race displacement, which is why application as a floating bearing is possible.

Bearings with ledges on the outer race are also used. If axial shaft fixation is required in the same direction, bearings with an extra ledge on the outer race are applied (Fig. 6.160b), and for axial fixing in two directions there are structures with an extra ledge and a thrust washer (Fig. 6.160c).

The double-row radial bearing with short cylindrical rollers and a conic hole (Fig. 6.160d) is used for high-speed shafts, which require precise rotation (mainly for spindles of metal-cutting machines). A high accuracy of the bearing in operation is reached on account of the structure workability, the possibility of adjusting the radial clearance by means of the inner race thrust, and the high rigidity caused by a large number of solids of revolution.

The roller radial spherical double-row bearing with ledges on the inner race (Fig. 6.160e) is intended for

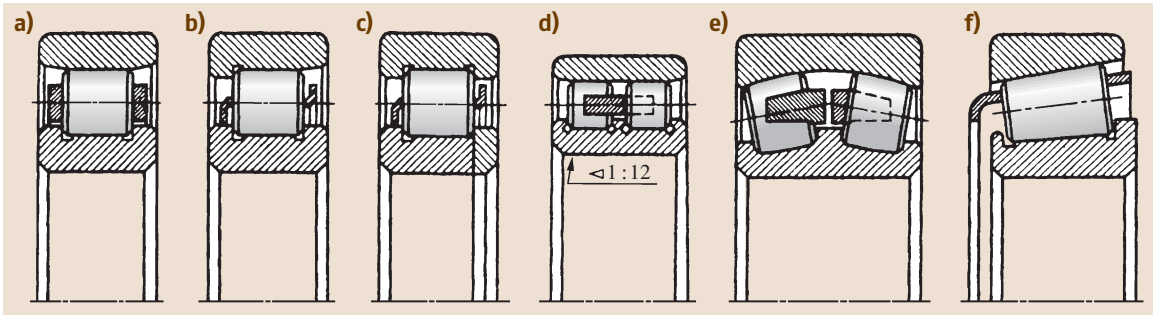


Fig. 6.160a–f Roller bearings. **(a)** Radial with short cylindrical rollers, **(b)** with shoulders on the inner race and an additional shoulder on the outer race, **(c)** with shoulders on the outer race and an additional shoulder on the inner race and a planar thrust washer, **(d)** radial double-row, **(e)** radial double-row spherical, **(f)** radial-thrust tapered

bearing radial and axial forces of any direction and allows considerable warp (0.5–2.5°) of the inner race (of the shaft) relative to the outer race (of the case). A rolling path of the outer race is made along the spherical surface. The rollers have the form of an asymmetrical or symmetrical barrel. This bearing differs from the ball radial spherical double-row bearing in its higher dynamic load rating, lower specific speed, and higher manufacturing complexity.

The roller radial-thrust tapered bearing with a contact angle of $\alpha = 10–16°$ (Fig. 6.160f) is intended for bearing radial and one-sided axial loads acting together.

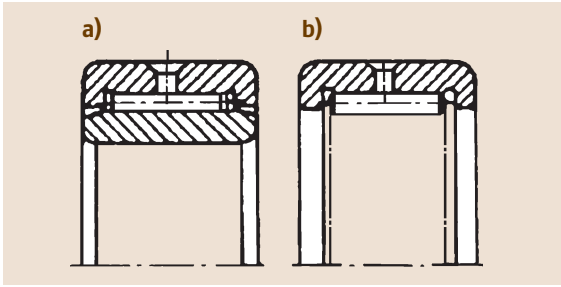


Fig. 6.161a,b Needle bearings. **(a)** With an outer and an inner race, **(b)** without inner race

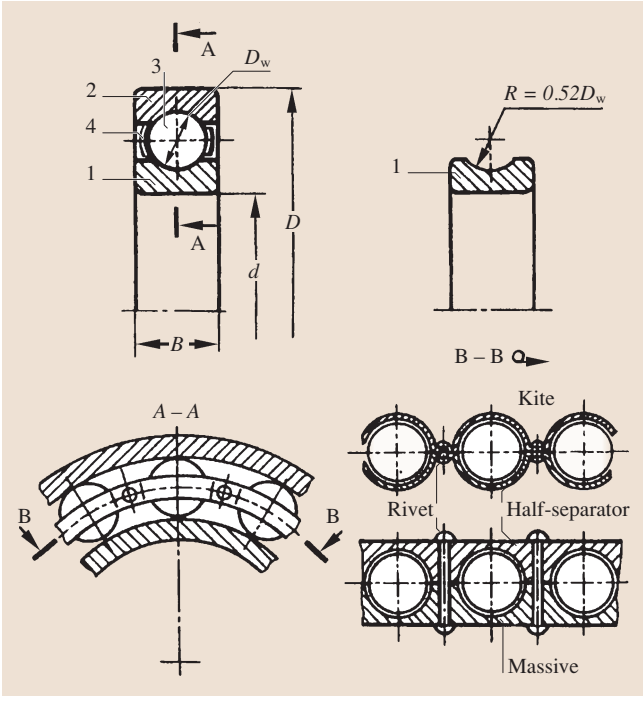


Fig. 6.162 The main components of a bearing

It differs from the ball radial-thrust bearing through its higher dynamic load rating, lower limit rotation frequency, and lower obtainable accuracy of the shaft rotation. The possibility of adjusting the axial bearing clearance must be considered in units with roller tapered bearings. Split assembly of the outer race and the inner race with a roller set is permissible.

This bearing has a rather wide application in mechanical engineering. It is remarkable for its convenient assembly, dismantling, and clearance adjustment. To provide pure rolling the tips of the tapered surfaces of the rolling paths of the races and the rollers must coincide. The rollers are aligned in the direction along the face from the side of the major diameter to avoid axial displacement. Bearings intended for particularly high axial loads are made with a contact angle of $\alpha = 20–30°$.

Tapered and cylindrical roller bearings are designed with a modified contact (rollers and rolling paths are made with a convex generatrix of different configuration), which means that the unfavorable effect of race axis warps can be decreased.

The roller radial needle bearing (Fig. 6.161a) is remarkable for its large radial dynamic load rating with small radial dimensions. It cannot bear axial forces and does not maintain axial shaft position.

Most of these structures are manufactured without cages. They are recommended for application in units running with hobbing motion or at low rotational frequencies. Needle bearings with cages can work with comparatively high rotational frequencies. They are sensitive to mutual warps of the races. Bearings with a profile modification allow insignificant mutual warps of the races.

To reduce size, a needle set is widely used in a cage without races or with one race (Fig. 6.161b). The surfaces of the shaft or the case for the needles must be quenched to high hardness, ground, and polished.

6.11.4 Functions of the Main Bearing Components

Figure 6.162 shows an axial section of a ball radial single-row bearing. The main components of the rolling bearing are 1 – the inner race, d – the hole diameter, 2 – the outer race, D – the outer diameter of the bearing, 3 – the solid of revolution – a ball, D_w – the diameter of the solid of revolution; 4 – the cage, which covers the solids of revolution and moves together with them.

The races of the bearings have grooves, which serve as guides for the solids of revolution.

A cage (see sections A–A and B–B in Fig. 6.162) is intended for direction, retention of the solids of revolution in a certain position (with the provision purpose of race coaxiality), and for separation of the solids of revolution from direct contact (with the objective of decreasing wear and friction losses). For low rotational frequencies and hobbing motion, bearings without cages are used (e.g., bearings of universal joint crosses).

The main application of a kite cage is in the configuration of two sinuous ring-shaped half-cages connected with rivets. Massive cages (unbroken or composite) are applied in fast-rotating units and high-accuracy bearings; they provide more precise positioning of the solids of revolution relative to the races of the bearings.

6.11.5 Materials of Bearing Components

In Russia races and solids of revolution are manufactured from special ball-bearing high-carbon chromium steels of various grades (100Cr6, 100CrMn6), as well as from carburized alloy steels (20MnCr5G, 17CrNiMO) (Appendix 6.A Table 6.95). The races have hardness 61–65 HRC, while the hardness of the solids of revolution is 63–67 HRC. Races and solids of revolution for bearings running at increased temperatures (up to 500 °C) or in corrosive mediums are produced from temperature-resistant or corrosion-resistant steels. For bearings that are to meet increased requirements in terms of lifetime and safety, steels subjected to special refining to decrease inclusion content, as well as double refining (electroslag and vacuum-arc refining), are used. In high-speed units bearings with balls made from silicon nitride Si_3N_4 are used.

In most cases cages are produced from soft carbon steel (Appendix 6.A Table 6.95). The rivets of the cages are manufactured from steel grades 15 and 20 (GOST) [6.89, 104, 106–115]. The cages of high-speed bearings are made massive from textolite, teflon, brass, and bronze, listed in ascending order of specific speed.

The surface roughness of the solids of revolution and rolling paths is $Ra = 0.04–0.08\ \mu m$.

Table 6.76 Designations of the inner diameters of bearings from 10 to 17 mm

Hole diameter (mm)	Designation
10	00
12	01
15	02
17	03

6.11.6 Nomenclature

The reference designation of a bearing is mostly marked on the race face. The main reference designation can consist of seven figures, indicating the diameter of the bearing hole, the dimensional series, the type, and the design philosophy. Zeros to the left of the least significant figure are not filled in. In this case, the number of figures in the reference designation is fewer than seven, e.g., 7208.

The first two figures to the right form a number that indicates the diameter, d, of the bearing hole. For bearings with $d = 20–495$ mm the hole diameter is determined by multiplication of this number by 5. Thus, bearing 7208 has $d = 40$ mm. Diameter designations of bearing holes from 10 to 17 mm are given in Table 6.76.

The third digit to the right indicates a diameter series and, together with the seventh digit designating a width series, determines the dimensional bearing series (Fig. 6.163). In ascending order of bearing outer diameter (with the same inner one) the series are designated as: 0, 8, 9, 1, 7, 2, 3, 4, 5, and 6. Thus, bearing 7208 has diameter series 2.

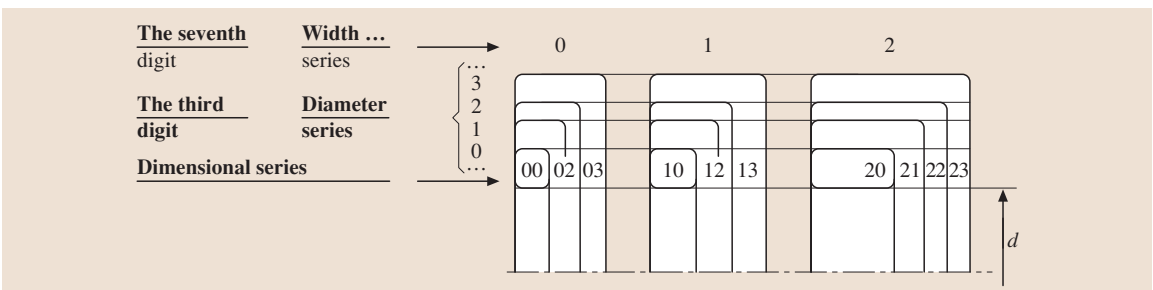


Fig. 6.163 Dimension series of bearings

The fourth figure to the right indicates a bearing type:

Table 6.77 Bearing type designation

Bearing type	Designation
Ball radial single-row	0
Ball radial spherical double-row	1
Roller radial with short cylindrical rollers	2
Roller radial spherical double-row	3
Roller needle or with long cylindrical rollers	4
Roller radial with spiral rollers	5
Ball radial-thrust single-row	6
Roller tapered	7
Ball thrust, ball thrust-radial	8
Roller thrust or thrust-radial	9

The example bearing 7208 is therefore roller tapered.

The fifth or the fifth and sixth digits to the right indicate a structural modification of the bearings (the value of the nominal contact angle in radial-thrust bearings, the presence of seals or a groove on the outer race for the thrust washer, etc.). The fifth and sixth digits in the bearing designation 7208 are absent; this is therefore the bearing with the main embodiment (base dimension type).

The seventh digit to the right indicates a width (height) series and with the third figure, which designates a diameter series, determines the dimensional series of the bearing. In ascending order of bearing width (height) the series are identified as: 7, 8, 9, 0, 1, 2, 3, 4, 5, and 6 (Fig. 6.163). The example roller-tapered bearing 7208 is of dimensional series 02.

Bearings of different types and series have dissimilar dimensions, mass m, dynamic load rating C_r, and limit rotation frequency $[n]$ (Fig. 6.164). In Fig. 6.164 most high-speed bearings are radial ball bearings from the dimensional series 02. The bearings in dimensional series 04 are not applicable for such high speeds, but their dynamic load rating is higher. Roller tapered bearings are characterized by a higher dynamic load rating than ball bearings with the same dimensions and lower limit rotation frequency.

Besides the digits of the main designation there may be additional alphabetic and numeric characters to the left or to the right, which define special manufacturing conditions of the given bearing.

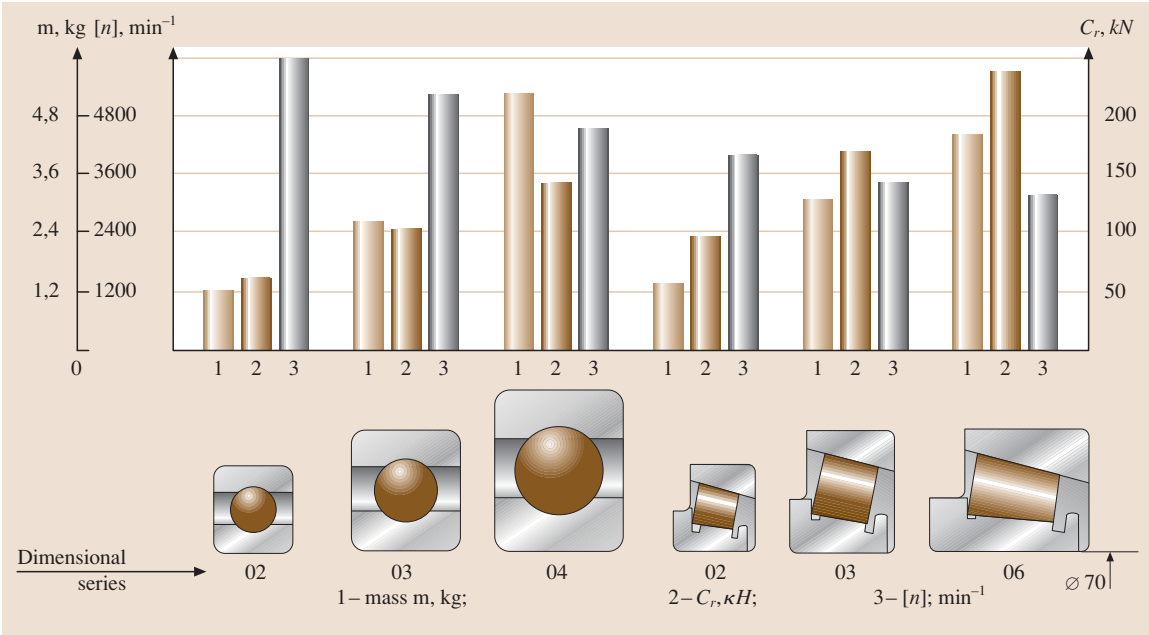


Fig. 6.164 Comparative parameters of radial ball and radial-thrust roller bearings of different dimension series

The characters defining an accuracy rating (0, normal, 6*X*, 6, 5, 4, *T*, 2, Appendix 6.B), a group for radial clearance (0, 1, 2–9; for radial-thrust ball bearings the grade of preinterference is indicated by 1, 2, and 3), a row for the frictional moment (1, 2–9) and a bearing class (A, B, and C) are marked to the left of the main designation.

The accuracy ratings are listed in ascending order of accuracy. In general engineering bearings of the accuracy ratings *normal* and 6 are used. In products with high accuracy or running with high rotational frequency (spindle units of high-speed machines, high-speed motors, etc.) bearings of classes 5 and 4 are applied. Bearings of accuracy rating 2 are used in gyroscopic devices.

The characters are located in recitation order right-to-left from the main designation of the bearing and are attached to it by a dash, e.g., A125-3000205, where 3000205 is the main designation, 5 is an accuracy rating, 2 is the group for radial clearance, 1 is a row of frictional moment; and A is the bearing class.

For all bearings except tapered ones the character "0" is used to designate the normal accuracy rating. For tapered bearings the character "0" is used to designate accuracy rating 0, the character "N" is for the normal accuracy rating, and the character "X" is for accuracy rating 6X. In our example the bearing 7208 has accuracy rating 0.

Depending on the presence of extra requirements for vibration level, deviations of shape and rolling surface position, frictional moment, etc., there are three bearing classes: A, increased regulated standards; B, regulated standards; and C, without extra requirements.

Possible characters to the right of the main designation are the following: A – increased dynamic load rating; E – the cage is made of plastic materials (polymers, textolite); P – the components of the bearing are made from heat-resistant steels; C1–C28 – closed class for filling with a lubricant; T (T1–T5) – temperature requirements of tempering of the bearing components, etc.

An example of the reference designation of a bearing with extra characters is A75-3180206ET2C2, i.e., a ball radial single-row bearing (0) with a double-sided seal (18) and a hole diameter of 30 mm (06), a diameter series 2, a width series 3, an accuracy rating 5, a radial clearance according to group 7, in the case of a requirement failure in the frictional moment, class A, with a cage made from plastic material (E), and the temperature of the regulating race tempering is 250 °C (T2), filled with a lubricant by the manufacturer (C2).

6.11.7 The Nature and Causes of Failure of Rolling Bearings

1. Fatigue flaking of the work surfaces of the races and solids of revolution in the form of bubbles or flaking-off under the action of fluctuating contact stresses. Nonmetallic inclusions in the steel, deep grinding marks, and microasperities are the main sources of crack nucleation. Fatigue flaking is the main fracture mode of bearings with good lubrication and ingress protection of the abrasive particles. It is usually observed after a long operation time.
2. Bearing of the work surfaces of the rolling paths and solids of revolution (formation of dimples and hollows) as a consequence of local plastic strains under the action of vibrational, impact, or considerable dead loads.
3. Abrasion owing to poor protection of the bearing from penetration of abrasive particles (construction-site engines, road and agricultural machines, looms). Application of perfect seal structures in bearing units decreases wear of the work-bearing surfaces.
4. Cage fracture due to the action of centrifugal forces and the influence of solids of revolution with different dimensions on the cage. This fracture mode is a principal cause of efficiency loss in high-speed bearings.
5. Fracture of the races and solids of revolution as a consequence of race warp and impact overloads (chipping of the ledges, splitting of the races, etc.). In the case of qualitative assembly and correct operation, component fracture of the bearings should not take place.

Outward signs of abnormal operation are the following: loss of rotational accuracy, increased noise and vibration, increased rotation, and temperature resistance. The main efficiency criteria for rolling bearings are contact fatigue strength and static contact strength.

6.11.8 Static Load Rating of Bearings

At initial point contact (ball bearings) touching of the bodies under load occurs along an elliptic area; at initial linear contact (roller bearings) it is along a rectangular area. The corresponding values of the contact stresses are determined from Hertz's formulas for point and linear contact.

The ratio of the curvature radii at the contact points is such that the contact stresses σ_h in the contact of the

solid of revolution with the inner race are higher than in the contact zone of the solid of revolution with the outer race for all bearing types (except spherical ones).

Thus, e.g., contact stresses σ_H (N/mm^2) for ball radial single-row bearings in contact with the inner race are

$$\sigma_H \approx 1035\sqrt[3]{F_0/D_w^2} \approx 1035\sqrt[3]{5F_r/\left(zD_w^2\right)}\,, \quad (6.22)$$

while for the outer race

$$\sigma_H \approx 827\sqrt[3]{F_0/D_w^2} \approx 827\sqrt[3]{5F_r/\left(zD_w^2\right)}\,,$$

where F_0 is a force acting on the most heavily loaded solid of revolution by the loading of the bearing with the radial force F_r (N), z is the number of solids of revolution, and D_w is the diameter of the solid of revolution (mm).

The basic static load rating of the bearing is a static load in N, which corresponds to the rated contact stress in the center of the most heavily stressed contact zone of the solid of revolution and the rolling path of the bearing.

According to the relevant ISO standard the following are assumed as design contact stresses σ_H for bearings:

Radial and radial-thrust ball (except self-installed):	$\sigma_H = 4200\,\mathrm{N/mm^2}$
Radial ball self-installed:	$\sigma_H = 4600\,\mathrm{N/mm^2}$
Radial and radial-thrust roller:	$\sigma_H = 4000\,\mathrm{N/mm^2}$
Thrust and thrust-radial ball:	$\sigma_H = 4200\,\mathrm{N/mm^2}$
Thrust and thrust-radial roller:	$\sigma_H = 4000\,\mathrm{N/mm^2}$

The total residual strain in the solid of revolution and the rolling path of the race arising by these contact stresses is approximately equal to 0.0001 of the diameter of the solid of revolution.

The static load rating for radial and radial-thrust bearings corresponds to the radial force F_r causing purely radial displacement of the races relative to each other. For thrust and thrust-radial bearings this corresponds to the central axial force F_a. The basic static load rating is designated in the following way: radial – C_{0r}, axial – C_{0a}.

With static loading, damage of the bearings appears in the form of the working surface plastic strain. In strength analysis the acting contact stress σ_H should be limited to

$$\sigma_H \le [\sigma]_H\,,$$

where $[\sigma]_H$ is the allowable contact stress.

The derivation of this formula is shown for the calculation of the basic static load rating using the example of a ball single-row radial bearing.

The strength condition for the most loaded point on the inner race of the bearing is, according to (6.22),

$$\sigma_H = 1035\sqrt[3]{5F_r/\left(zD_w^2\right)} \le [\sigma]_H\,.$$

From this the allowable radial load is

$$[F]_r = \left[\frac{1}{5}\left(\frac{[\sigma]_H}{1035}\right)^3\right] zD_w^2\,.$$

Designating the expression in square brackets on the right-hand side by f_0 and writing $[F]_r = C_{0r}$ we obtain the formula to calculate the basic static load rating C_{0r} (N), for radial and radial-thrust ball bearings:

$$C_{0r} = f_0 i z D_w^2 \cos\alpha\,,$$

where f_0 is a coefficient depending on the bearing class, material, and geometry of the bearing components, their manufacturing accuracy, and the assumed value of the design contact stress (Table 6.78); i is the number of rows of the solids of revolution, z is the number of solids of revolution in a row, D_w is the ball diameter (mm), and α is the nominal contact angle (degrees).

Design dependencies for the calculation of the static load rating for other bearing classes are given in the standard [6.111].

The values of the basic static load rating C_{0r} (C_{0a}) for all bearings are calculated in advance and given in the manufacturer's catalog.

6.11.9 Lifetime Testing of Rolling Bearings

The lifetime is the running time of the bearing until the appearance of signs of material fatigue on the solids of revolution or the races. The bearing lifetime is designated by L (life) and is expressed in terms of the number of millions of revolutions of one race relative to another or in terms of working hours. The main design dependencies for matching of bearings are obtained on the basis of a pilot study of specimen and full-scale bearings.

Figure 6.165 shows a contact stress-cycle diagram of specimens manufactured according to the standards of the bearing industry technology. The ordinate axis shows contact stresses σ_H, which were determined according to Hertz's theory; the abscissa shows the lifetime, expressed by the number N of stress change cycles to fracture. Stress-cycle diagrams are plotted for the different probability levels Q of fracture: 0.01, 0.10, 0.30,

Table 6.78 Values of the coefficient f_0 for ball bearings. The values f_0 are determined from Hertz's formulas obtained from the condition of initial point contact with a modulus of elasticity of 2.07×10^5 N/mm^2 and Poisson's ratio of 0.3. The values of f_0 are calculated for the case of common external force distribution between the solids of revolution, when the load on the most loaded ball in ball radial and radial-thrust bearings is equal to $5F_r/(z \cos\alpha)$, and in ball thrust and thrust-radial bearings it is $F_a/(z \sin\alpha)$. f_0 for the intermediate values $D_w \cos\alpha / D_{pw}$ is calculated by linear interpolation

$D_w \cos\alpha / D_{pw}$	f_0 for ball bearings		
	Radial and radial-thrust	Self-installed	Thrust and thrust-radial
0.00	14.7	1.9	61.6
0.01	14.9	2.0	60.8
0.02	15.1	2.0	59.9
0.03	15.3	2.1	59.1
0.04	15.5	2.1	58.3
0.05	15.7	2.1	57.5
0.06	15.9	2.2	56.7
0.07	16.1	2.2	55.9
0.08	16.3	2.3	55.1
0.09	16.5	2.3	54.3
0.10	16.4	2.4	53.5
0.11	16.1	2.4	52.7
0.12	15.9	2.4	51.9
0.13	15.6	2.5	51.2
0.14	15.4	2.5	50.4
0.15	15.2	2.6	49.6
0.16	14.9	2.6	48.8
0.17	14.7	2.7	48.0
0.18	14.4	2.7	47.3
0.19	14.2	2.8	46.5
0.20	14.0	2.8	45.7
0.21	13.7	2.8	45.0
0.22	13.5	2.9	44.2
0.23	13.2	2.9	43.5
0.24	13.0	3.0	42.7
0.25	12.8	3.0	41.9
0.26	12.5	3.1	41.2
0.27	12.3	3.1	40.5
0.28	12.1	3.2	39.7
0.29	11.8	3.2	39.0
0.30	11.6	3.3	38.2
0.31	11.4	3.3	37.5
0.32	11.2	3.4	36.8
0.33	10.9	3.4	36.0
0.34	10.7	3.5	35.3
0.35	10.5	3.5	34.6
0.36	10.3	3.6	–
0.37	10.0	3.6	–
0.38	9.8	3.7	–
0.39	9.6	3.8	–
0.40	9.4	3.8	–

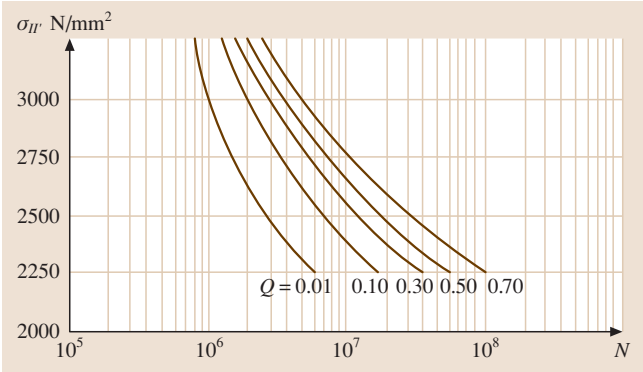


Fig. 6.165 Contact-stress cycle diagram

0.50, and 0.70. The equation for the curve in the stress-cycle diagram is $\sigma_{Hi}^q N_i = \text{const}$. The exponent is $q = 9$ for point contact and $q = 20/3$ for linear contact.

According to the test results of 20–30 specimens the lifetime exhibits considerable dispersion at the same stress level. The maximum value can differ from the minimum value by 50–100.

Dispersion of the testing results occurs as a consequence of the static nature of the fatigue failure process, which is caused by microstructure discontinuities in the metal: different dimensions, forms, and metal grain orientation, the presence of different structural phases, nonmetallic inclusions, different orientation of the crystal lattice, as well as occasional changes in microgeometry and the structure of the surface layer, etc.

The load P on the specimen or the bearing corresponds to the contact stresses σ_H. For a point contact

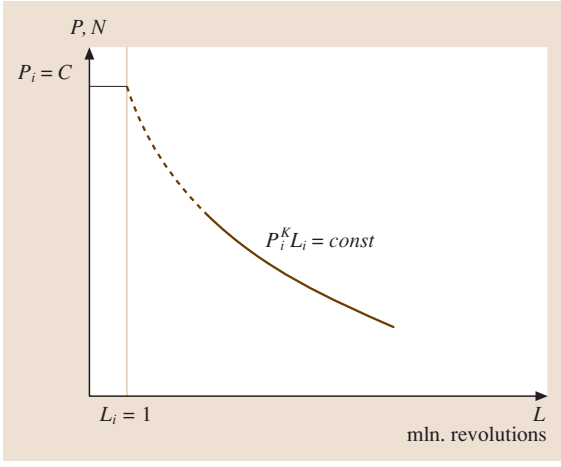


Fig. 6.166 Dependence of the bearing lifetime L on the acting load P

P is proportionate to σ_H^3, for a linear contact $P \sim \sigma_H^2$. On the basis of pilot studies the following dependence is fixed between P acting on the bearing load and its lifetime L_i (Fig. 6.166)

$$P_i^k L_i = \text{const.}\,, \tag{6.23}$$

where k is the exponent of the stress-cycle diagram ($k = 3$ for ball bearings and $k = 10/3$ for roller bearings).

Theoretically, the stress-cycle diagram, which is obtained experimentally, can be extrapolated to the area of higher loads – hatch in Fig. 6.166. Assuming the lifetime of the bearing to be $L_i = 1$ million revolutions and indicating the load P_i, which corresponds to this lifetime, through C, it can be written in accordance with (6.23) for the stress-cycle diagram as

$$P_i^k L_i = C^k 1\,.$$

Omitting the index i, we obtain formula for calculation of the lifetime L, in millions of revolutions, depending on the bearing load P (N), in a general form as

$$L = (C/P)^k\,. \tag{6.24}$$

The load C (N) is called the *dynamic load rating*. The design dependence (6.24)) is correct when $P \le 0.5C$.

In view of the considerable dispersion in the fatigue characteristics, the testing results are handled with a statistical method. The distribution of bearing failure is described according to Weibull's two- or three-parameter distribution.

Based on estimation of the test results the value of the lifetime L_{10} that corresponds to the probability $Q = 10\%$ of bearing fatigue failure is used.

6.11.10 Design Dynamic Load Rating of Bearings

The basic dynamic radial (or axial) design load rating represents a radial (or axial) load in N, which the rolling bearing can theoretically support for the base design lifetime constituting 1 million revolutions.

The basic design lifetime L_{10} is the lifetime in millions of revolutions that corresponds to 90% safety for a specific bearing or a group of identical rolling bearings manufactured from a common material with the application of standard technology and running under the same standard operating conditions, i.e., the bearing is installed correctly, lubricated, protected from the penetration of foreign bodies, and the load corresponds to the bearing dimension type, and the bearing is not

subjected to excessive changes of temperature and rotational frequency.

The basic dynamic design load rating is designated in the following way: radial – C_r, axial – C_a. The values C_r (C_a) for each bearing are calculated in advance and indicated in the manufacturer's catalog. The development of the formula for calculation of the basic dynamic radial load rating is shown by using the example of a ball radial single-row bearing.

The calculation is based on the use of experimental stress-cycle diagrams (Fig. 6.165) described by a dependence

$$\sigma_{Hi}^{q} N_i = \text{const.} , \tag{6.25}$$

where $q = 9$ for ball bearings, and const. is a constant that corresponds to the experimental environment.

With bearing loading with a radial force F_r and inner race rotation and a nonrotating outer race, the number of loading cycles for L million revolutions is

$$N = 0.5 \times 10^6 z K_1 K_{ef} L ,$$

where z is the number of solids of revolution, $K_{ef} < 1$ is an equivalence coefficient taking into account the uneven load distribution between the solids of revolution, and $K_1 = 0.5(D_{pw} + D_w \cos\alpha)/D_{pw}$. Here D_{pw} is the diameter of the circle going through the centers of the solids of revolution, D_w is the ball diameter, and α is the contact angle.

In accordance with (6.25) we have

$$\left[1035 \sqrt[3]{5F_r/\left(zD_w^2\right)}\right]^9 0.5 \times 10^6 z K_1 K_{ef} L = \text{const.}$$

The left- and the right-hand sides of this expression are raised to a power of one-third and, after transformation, we have

$$F_r L^{1/3} = \left[\frac{\text{const}^{1/3}}{1035^3 \times 5 \left(0.5 \times 10^6 K_1 K_{ef}\right)^{1/3}}\right] \times z^{2/3} D_w^2 .$$

The expression in square brackets is designated f_c. In accordance with (6.24) for $P = F_r$ we have $F_r L^{1/3} = C_r$. After appropriate changes and some corrections we obtain the formula for the calculation of C_r (N), the basic dynamic radial design load rating for ball radial and radial-thrust bearings

$$C_r = b_m f_c (i \cos\alpha)^{0.7} z^{2/3} D_w^{1.8} ,$$
$$\text{by} \quad D_w \le 25.4\,\text{mm} ;$$
$$C_r = 3.647 b_m f_c (i \cos\alpha)^{0.7} z^{2/3} D_w^{1.4} ,$$
$$\text{by} \quad D_w > 25.4\,\text{mm} ,$$

where b_m is a coefficient characterizing the behavior of steel, taking into account its method of manufacture and depending on the bearing class and structure; f_c is a coefficient depending on the geometry of the bearing components and their production accuracy; i is the number of rows of solids of revolution; and z is the number of solids of revolution in a row.

Design dependencies for the calculation of the dynamic load rating are given in standards for other bearing classes. By definition, the basic dynamic load rating represents a very large load corresponding to the theoretical area of the stress-cycle diagram, that is not achievable in practice.

6.11.11 Design Lifetime of Bearings

The basic design lifetime L_{10} in millions of revolutions is determined by 90% safety (as indicated by the figure 10 in the designation; i.e., $10 = 100 - 90$)

$$L_{10} = \left(\frac{C}{P}\right)^k , \tag{6.26}$$

where C is the base dynamic load rating of the bearing (radial C_r or axial C_a) (N), P is the equivalent dynamic load (radial P_r or axial P_a) (N), and k is an exponent that is chosen in accordance to the outcomes of experiments to be $k = 3$ for ball bearings and $k = 10/3$ for the roller bearings.

The formula for the lifetime calculation is correct if P_r (or P_a), and for varying loads $P_{r\,max}$ (or $P_{a\,max}$), does not exceed $0.5C_r$ (or $0.5C_a$). The applicability of this formula is also limited to rotational frequencies from $10\,\text{min}^{-1}$ to the limiting values stated in the manufacturer's catalog.

From the given formula the basic design lifetime L_{10} for bearings that are produced from standard bearing steels according to standard technology and exploited under common conditions is calculated.

For different material properties or operating conditions from the standard ones, as well as for increased safety requirements and to take into account special bearing properties, the *corrected design lifetime* L_{sa} is determined in millions of revolutions as

$$L_{sa} = a_1 a_2 a_3 L_{10} , \tag{6.27}$$

where a_1 is a coefficient correcting the lifetime depending on the *safety* P_t (Table 6.41), a_2 is a coefficient adjusting the lifetime depending on the special bearing properties, and a_3 is a coefficient correcting the lifetime depending on the operating conditions of the bearing.

The corrected design lifetime of the bearing in operating hours is

$$L_{sah} = 10^6 L_{sa}/(60n) ,$$

where n is the rotational frequency of the race (min^{-1}).

Sometimes it is more convenient to express the bearing lifetime of vehicles (the bearings of wheel hubs and half-axes) in units of distance. The corrected design lifetime in millions of kilometers is

$$L_{sas} = (\pi D/1000) L_{sa} ,$$

where D is the wheel diameter in meters.

The rolling bearing calculation for an increased probability of nonfailure during operation is carried out for important units by using a safety factor of 91–99%. Instead of the index s, the value of the difference $(100 - P_t)$ is written in the lifetime designation, where P_t is the safety used for the lifetime determination. Thus, for 90% safety one writes L_{10a} (L_{10ah}) and for 97% safety one writes L_{3a} (L_{3ah}).

The bearing can obtain special properties, resulting in a different lifetime, following the application of special materials (e.g., steels with a particularly low content of nonmetallic inclusions) or special production processes, or are of a special structure. The values of the coefficient a_2 are fixed by the bearing manufacturer.

Working conditions, which are additionally taken into account with the help of the coefficient a_3 – it is a conformity of the lubricant viscosity with the required value (taking into account rotational frequencies and temperatures), the presence of foreign particles in the lubricant, as well as conditions causing material property change of the bearing components (e.g., high temperature causes hardness to decrease).

Calculation of the basic lifetime is built upon the fact that the thickness of the oil film in the contact zones rolling element race is equal to or a little more than the total roughness of the contact surfaces; Therefore $a_3 = 1$. The bearing producer provides recommendations concerning the values of the coefficient a_3 for other conditions.

For the choice of the bearing dimension and the calculation of the corrected lifetime for specific operating conditions it is supposed that the bearings correspond to the required accuracy grade and that the required strength and rigidity of the shafts and cases are provided.

Application of the values $a_2 > 1$ and $a_3 > 1$ in the formula of the corrected lifetime will be valid.

6.11.12 The Choice of Bearing Classes and Their Installation Diagrams

Each bearing class has particular features due to its structure. For the choice of bearing class a few different comparative factors must be appreciated. Thus, it is impossible to formulate a general law for bearing choice. The most significant factors are given below:

- Value and direction of the load (radial, axial, combined)
- Load conditions (constant, varying, vibrational, impact)
- Rotational frequency of the bearing race
- Required lifetime (in hours or millions of revolutions)
- Environmental conditions (temperature, humidity, dust level, acidity, etc.)
- Particular requirements for the bearing, which are made with a unit structure (the necessity for bearing self-installation into the support for warp compensation of the shaft or the case; the ability to allow shaft displacement in the axial direction; bearing assembly directly onto the shaft, on the clamping or clamping-tightening sleeve; the necessity to adjust the radial and axial clearance of the bearing, increase of support rigidity and accuracy of the shaft rotation, decrease of the frictional moment, noisiness; desired overall dimensions of the unit, requirements for safety; price of the bearing and of the entire unit)

With the choice of bearing type the common practice in machine design and operation of the fixed machine class can be headed for. Thus, for example, ball radial bearings are mostly used for shaft supports of spurs and helical wheels, reduction gears, and gearboxes. Tapered roller bearings are applied as shaft supports of spur gears where the dimensions of the ball bearings are excessively large.

Bevel and worm wheels must be precisely and rigidly fixed in the axial direction. Ball radial bearings are characterized by low axial rigidity. Thus, in power trains, tapered roller bearings are used for shaft supports of bevel and worm wheels.

For the shaft supports of the bevel pinion tapered roller bearings are applied from the same considerations. For high rotational frequency of the gear shaft ($n > 1500\,\mathrm{min}^{-1}$) ball radial-thrust bearings are used.

Worm supports in power worm gears are loaded with considerable axial forces, which is why tapered roller bearings are mainly applied as supports for worm

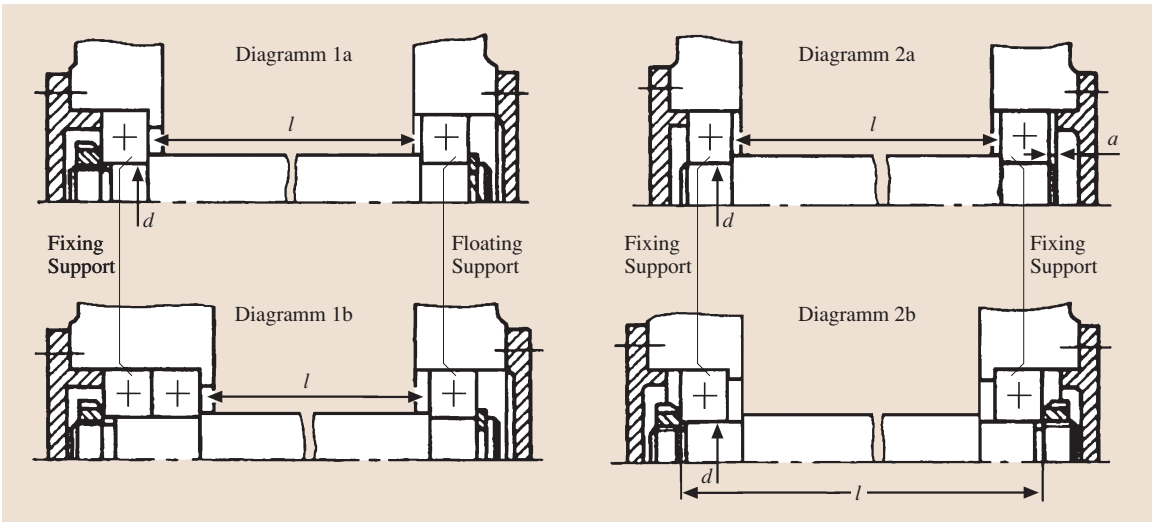


Fig. 6.167 Methods of axial fixation of the shafts

shafts. Under continuous operation of the worm gear ball radial-thrust bearings are also used in order to reduce heat generation.

For the supports of floating shafts of herring-bone gears, radial bearings with short cylindrical rollers are used.

Bearings of the normal accuracy degree are usually used. Bearings of higher accuracy are applied for shaft supports that require increased rotation accuracy or run at particularly high rotational frequencies. Application of bearings with higher accuracy degrees increases product value.

After class, structural modifications and the installation diagram of the bearings are outlined, the bearing is chosen from the catalog, and the lifetime calculation and/or static load rating calculation are performed for the required safety level. Depending on the operating rates and working conditions, the lubrication method, lubricant type, its protection from contamination, and outflow from the bearing are then chosen.

Installation diagrams

In most cases the shafts should be fixed to the supports to avoid axial displacements. The shafts are divided into fixing and floating types according to their ability to fix the axial position of the shaft. Axial shaft displacement is restricted in fixed supports in one or both directions. In a floating support axial shaft displacement is not limited in any direction. A fixing support is loaded only with radial and axial forces, whereas a floating support is only loaded with radial forces.

In some structures so-called *floating* shafts are used. These shafts allow axial displacement in both directions; they are installed on the floating supports. Fixation of the axial shaft is not carried out in the supports, but by some other structural components, e.g., with the component faces or the teeth of chevron gears.

Figure 6.167 shows the main methods of axial shaft fixation. In diagrams 1a and 1b (Fig. 6.167) the shaft is fixed in a (left-hand side of the figure) support: in 1a a bearing is used, whereas in 1b two single-row bearings are used. Radial bearings are usually applied in a floating support. These diagrams can be used for any distance l between the shaft supports. The diagramm 1b (Fig. 6.167) is characterized by high rigidity of the fixing support, especially in the case of the application of two radial-thrust bearings with large contact angles in one support.

By combining a fixing support and a floating support, the aim is to provide approximately even loading of the bearing and the lowest frictional forces in the floating support.

With fluctuations in temperature the floating bearing moves in the axial direction due to the extension (shortening) of the shaft. Then displacement can take place under load, and the hole surface of the case wears out. This is why, under the action of only radial forces on the shaft supports, the less loaded support is chosen as the floating one. If the output shaft end is connected to the shaft of another unit by means of a sleeve, the support at the output shaft end is assumed to be fixed.

In diagrams 2a and b (Fig. 6.167) the shaft is fixed in two supports, with each support in one direction. These diagrams are used with certain limitations on the distance between the supports, which results from the change in clearance at the bearings as a consequence of component heating during operation. With heating of the bearings the clearances in them decrease; with shaft heating its length increases.

Due to the increase of the shaft length the axial clearances in the bearings of diagram 2a (Fig. 6.167) are also reduced. The axial clearance a is designed for the supports at assembly so that shaft jamming does not occur. The value of the clearance must be slightly higher than the expected thermal deformation of the bearings and the shaft. Depending on the unit structure and operating conditions one uses $a = 0.15–1.0$ mm.

Installation diagram 2a (Fig. 6.167) is physically the simplest. It is widely used with relatively short shafts. On assembly of the radial bearings in the supports the ratio is $l/d \approx 8–10$, and no higher.

In the supports in diagram 2a (Fig. 6.167) radial-thrust bearings can also be used. As these bearings are more sensitive to changes in axial clearances, more exacting requirements are made on the ratio l/d, i. e., $l/d \leq 6–8$. Lower values are used for roller bearings, whereas higher values are used for ball radial-thrust bearings. It is not recommended to apply radial-thrust bearings with a contact angle of $\alpha = 25 - 40°$ according to this diagram.

With the shaft installation according to diagram 2b (Fig. 6.167) the probability of the bearings jamming as a consequence of thermal deformations is lower, because the axial clearance in the bearings increases with shaft length extension. The distance between the bearings can be larger in diagram 2a (Fig. 6.167): for ball radial bearings $l/d = 10–12$, for ball radial-thrust bearings $l/d \leq 10$, and for tapered roller bearings $l/d \leq 8$.

It is not recommended to install longer shafts according to diagram 2b (Fig. 6.167), as large axial clearances can occur due to thermal deformations, which are inadmissible for radial-thrust bearings.

6.11.13 Determination of Forces Loading Bearings

Determination of Radial Reactions

The shaft on the bearings installed one at a time in the support is conditionally considered as a bar on hinged movable supports or as a bar with one hinged movable support and one hinged immovable support. The radial reaction F_r of the bearing is regarded as being applied to the shaft axis at the crossing point of the lines normal to it, drawn through the centers of the contact areas. For radial bearings this point is located in the center of the bearing width. For radial-thrust bearings the distance a between this point and the bearing face can be determined graphically (Fig. 6.168) or analytically as:

- For ball bearings: $a = 0.5[B + 0.5(d + D)\tan\alpha]$
- For tapered roller bearings: $a = 0.5[T + (d + D)e/3]$

where B is the bearing width, T is the mounting height, d is the hole diameter, D is the outer diameter, α is the nominal contact angle, and e is the coefficient of axial loading. The values of the listed parameters are taken from the manufacturer's catalog.

Supporting forces are determined from the equilibrium equation: the total moment of external forces relative to the relevant support and the reaction moment in another support is equal to zero.

In a number of cases, the rotational direction can vary or be uncertain, and the change in rotational direction can result in modification of not only the direction but also the values of the supporting forces. With installation of a joint sleeve on the shaft ends the direction of the force on the shaft due to the the sleeve is unknown. In such cases, the most dangerous situation is considered by the reaction calculation. In order to create favorable working conditions for ball and roller bearings, they must be subject to a certain minimum continuously. This is especially so if they run at high rotational frequencies, in which case the inertial forces of

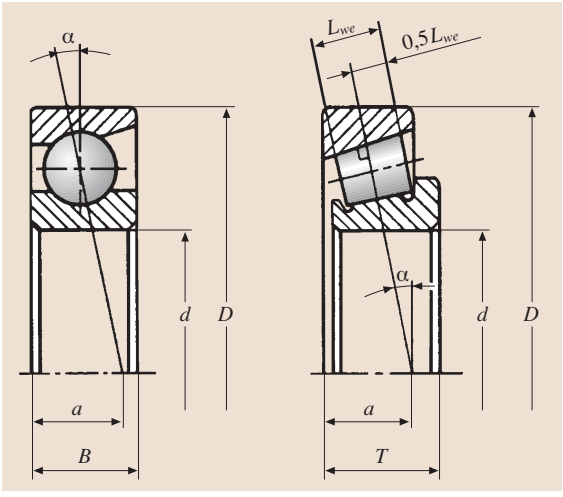


Fig. 6.168 Position of the application point of the radial reaction in radial-thrust bearings

Table 6.79 Formulas for the calculation of the coefficients X, Y, and e for ball radial and radial-thrust bearings. For single-row bearings with $F_a/F_r \le e$ it is assumed that $X = 1$ and $Y = 0$. In the formulas given in the table C_{0r} is a static load rating of the bearing; for the double-row bearings C_{0r} is a static load rating of a row (half of the static load rating of the double-row bearing)

Bearing class	α (°)	Axial loading coefficient e	Single-row bearing		Double-row bearing			
			$F_a/F_r > e$		$F_a/F_r \le e$		$F_a/F_r > e$	
			X	Y	X	Y	X	Y
Radial	0	$0.28\left(\frac{f_0 F_a}{C_{0r}}\right)^{0.23}$	0.56	$0.44/e$	1.0	0	0.56	$0.44/e$
Radial-thrust	12	$0.41\left(\frac{f_0 F_a}{C_{0r}}\right)^{0.17}$	0.45	$0.55/e$	1.0	$0.62/e$	0.74	$0.88/e$
	15	$0.46\left(\frac{f_0 F_a}{C_{0r}}\right)^{0.11}$	0.44	$0.56/e$	1.0	$0.63/e$	0.72	$0.91/e$
	18	0.57	0.43	1.0	1.0	1.09	0.70	1.63
	25 26	0.68	0.41	0.87	1.0	0.92	0.67	1.41
	36	0.95	0.37	0.66	1.0	0.66	0.60	1.07
	40	1.14	0.35	0.57	1.0	0.55	0.57	0.93

the solids of revolution and the cage, as well as friction in the lubricant, can negatively influence rolling conditions in the bearing and cause creep of the balls and rollers along the rolling path.

As a general recommendation it is assumed that the loads that affect the roller bearings must be $0.02C$ and those on the ball bearings must be $0.01C$, where C is the dynamic load rating. The weight of the components supported by the bearing with the external forces often exceeds the required minimum load. Otherwise the bearing must be loaded by an extra radial or axial force. It is easier to provide such a force, e.g., in systems with radial and radial-thrust ball bearings, and tapered roller bearings, by means of prior axial loading made by adjustment of the relative position of the inner and outer races with spacing racers, pads, or springs. Extra radial force can be applied in the same way, e.g., by means of increased belt tension.

Determination of Axial Reactions

By the installation of a shaft on two nonadjustable radial ball or radial-thrust bearings the axial force F_a loading the bearing is equal to the external axial force F_A acting on the shaft. The force F_A supports the bearing, which limits the axial displacement of the shaft under the action of this force.

By determination of the axial forces loading adjustable radial-thrust bearings the axial forces that arise under the action of the radial load F_r as a consequence of the tilt of the contact area with respect to the hole axis of the bearing should be taken into account. The values of these forces depend on the bearing class, contact angle, and the radial forces, as well as on how the bearings are adjusted. If the bearings are assembled with a large clearance, only one or two balls or rollers take the whole load. The axial load component equals $F_r \tan\alpha$ for transmission through only one solid of revolution. The working conditions of the bearings are unfavorable with large clearances, so such clearances are not permissible. Bearings are usually adjusted in such a way that the axial clearance is about zero under fixed temperature conditions. In this case, about half of the solids of revolution are under the action of the radial load F_r, and the total axial component for all the loaded

Table 6.80 Values of the coefficients X, Y, and e for roller radial-thrust bearings ($\alpha \neq 0°$)

Bearing classes	X	Y	X	Y	e
	$F_a/F_r \le e$		$F_a/F_r > e$		
Single-row	1.0	0	0.4	$0.4\cot\alpha$	$1.5\tan\alpha$
Double-row	1.0	$0.45\cot\alpha$	0.67	$0.67\cot\alpha$	$1.5\tan\alpha$

solids of revolution is equal to $e' F_r$ as a consequence of the tilt of the contact area and represents a minimum axial force, which must act on the radial-thrust bearing with the set radial force

$$F_{amin} = e' F_r \,, \tag{6.28}$$

where e' is a coefficient of the minimum axial load.

For ball radial-thrust bearings with contact angle $\alpha < 18°$, one has $F_{amin} = e' F_r$. In such bearings the actual contact angle differs from the initial (nominal) one and depends on the radial load F_r and basic static load rating C_{0r}. Thus, the coefficient e' is determined depending on the ratio F_r/C_{0r} from the following formulas:

- For bearings with contact angle $\alpha = 12°$

$$e' = 0.57(F_r/C_{0r})^{0.22} \,.$$

- For bearings with contact angle $\alpha = 15°$

$$e' = 0.58(F_r/C_{0r})^{0.14} \,.$$

- For ball radial-thrust bearings with contact angle $\alpha \geq 18° e' = e$ and $F_{amin} = eF_r$. The values of the coefficient e for axial loading are taken from Table 6.79.
- For tapered roller bearings $e' = 0.83e$ and $F_{amin} = 0.83eF_r$. The values of the coefficient e are taken from Table 6.80 or from the manufacturer's catalog.

In order for normal working conditions to be guaranteed, the axial force loading the bearing must not be lower than the minimum one, $F_a \geq F_{amin}$. This condition must be met for every support.

If $F_a \geq F_{amin}$, more than half or all of the bearing solids of revolution are under load. The support rigidity increases with increasing axial load, which is why in some supports, e.g., for work spindles, assembly with prior interference is applied.

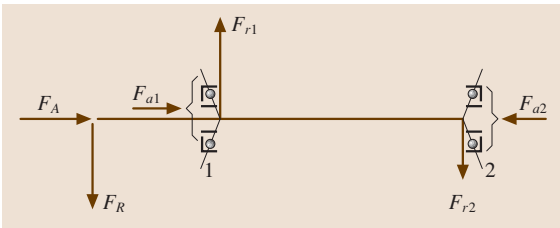


Fig. 6.169 Loading diagram of the shaft and supports with radial-thrust adjustable bearings

Under normal operation of radial-thrust bearings it is necessary that the axial force loading the bearing is above the minimum force in all supports

$$F_{a1} \geq F_{a1\,min} \quad \text{and} \quad F_{a2} \geq F_{a2\,min} \,.$$

Moreover, the equilibrium condition of the shaft must be met, i.e., the sum of all the axial forces acting on the shaft must be zero; for example, for the configuration shown in Fig. 6.169 we have

$$F_A + F_{a1} - F_{a2} = 0 \,.$$

Example of the Determination of Axial Supporting Forces

In the design model represented in Fig. 6.169 it is indicated that F_A and F_R are external axial and radial loads acting on the shaft, F_{r1} and F_{r2} are radial supporting forces, and F_{a1} and F_{a2} are axial supporting forces.

The solution can be obtained from joint satisfaction of three equations:

- From the condition that $F_a \geq F_{amin}$ in every support, and taking into account (6.28), it follows that:

$$F_{a1} \geq e'_1 F_{r1}, \; F_{a2} \geq e'_2 F_{r2} \,.$$

- From the equilibrium condition of the shaft under the action of the axial forces it follows that:

$$F_A + F_{a1} - F_{a2} = 0 \,.$$

The method of attempts is applied, previously assuming that the axial force in one of the supports is equal to the minimum value.

1. Let, for example, $F_{a1} = e'_1 F_{r1}$. Then from the equilibrium condition of the shaft we have

$$F_{a2} = F_A + F_{a1} = F_A + e'_1 F_{r1} \,.$$

   Check the condition fulfillment $F_a \geq F_{amin}$ for the second support. If $F_{a2} \geq e'_2 F_{r2}$, the axial forces F_{a1} and F_{a2} have been determined correctly. If $F_{a2} < e'_2 F_{r2}$ (which is inadmissible), the initial assumption was false. A second attempt should be made.
2. It should be assumed that $F_{a2} = e'_2 F_{r2}$. Then from the equilibrium condition of the shaft we have

$$F_{a1} = F_{a2} - F_A = e'_2 F_{r2} - F_A \,.$$

The condition $F_{a1} \geq e'_1 F_{r1}$ will then certainly be met.

6.11.14 Choice and Calculation of Rolling Bearings

Let us assume that the type and configuration of the bearing installation have been chosen previously. The dimensions of the bearing selected for the application can be chosen on the basis of estimation of its loading rate in accordance with the corresponding acting loads, rotational frequency, required lifetime, and safety. The values of the dynamic and static load ratings are given in the manufacturer's catalog. Calculations on the static and/or dynamic load ratings are now to be performed.

The static load rating is not only used to calculate the parameters for nonrotating bearings or those rotating with low rotational frequencies ($n < 10\,\mathrm{min}^{-1}$), or those performing slow oscillatory rotations, but also for bearings rotating with frequency $n \geq 10\,\mathrm{min}^{-1}$ and those subjected to the action of short-term impact loads or substantial overloads. The static load rating of bearings that run with low rotational frequencies and are designed for a short lifetime are also checked in this way.

Calculations *of the dynamic load rating* (specified lifetime calculation) are performed for the entire load range. Testing is additionally carried out under the assumption of the application of the highest loads.

Calculation of the Static Load Rating of Bearings

A static load rating calculation checks whether the equivalent static load P_{0r} (P_{0a}) on the bearing exceeds the static load rating C_{0r} (C_{0a}) given in the manufacturer's catalog

$$P_{0r} \leq C_{0r} \quad \text{or} \quad P_{0a} \leq C_{0a} \,.$$

The equivalent static radial (or axial P_{0a}) load P_{0r} is a static radial (or axial) load that causes the same contact stress in the most heavily loaded contact zone as in the conditions of actual loading.

The static equivalent radial load for ball radial and radial-thrust, and roller radial-thrust ($\alpha \neq 0°$) bearings is equal to the greater of the two values determined from the expressions

$$P_{0r} = X_0 F_r + Y_0 F_a \,;$$
$$P_{0r} = F_r \,,$$

where F_r and F_a are, respectively, the radial and axial loads on the bearing (N), and X_0 and Y_0 are, respectively, the coefficients of the static radial and static axial loads (Table 6.81).

For roller radial bearings $\alpha = 0°$, which supports only a radial load, $P_{0r} = F_r$. The static equivalent axial load for ball and roller thrust-radial bearings ($\alpha \neq 90°$) is determined from

$$P_{0a} = 2.3 F_r \tan\alpha + F_a \,.$$

For ball and roller thrust bearings ($\alpha = 90°$) one has $P_{0a} = F_a$. For the calculation of the static equivalent radial load for two identical single-row radial ball, radial-thrust ball, and roller bearings installed together on the same shaft positioned with the wide or narrow faces towards one other, making a mutual bearing unit, the values X_0 and Y_0 for double-row bearings are used, and the values F_r and F_a are assumed to form a combined load acting on the whole set.

For the choice and calculation of the bearings it should be borne in mind that the allowable static equivalent load P_0 can be lower than, equal to, or higher than the basic static load rating. The value of this load depends on the requirements of run smoothness (e.g., of

Table 6.81 Values of the coefficients X_0 and Y_0. The values Y_0 for the intermediate contact angles are obtained by means of linear interpolation

Bearing class		Single-row bearings		Double-row bearings	
		X_0	Y_0	X_0	Y_0
Ball radial		0.6	0.5	0.6	0.5
Ball radial-thrust with contact angle α (°)	12	0.5	0.47	1.0	0.94
	15		0.46		0.92
	20		0.42		0.84
	25		0.38		0.76
	30		0.33		0.66
	35		0.29		0.58
	40		0.26		0.52
	45		0.22		0.44
Ball and roller self-installed, $\alpha \neq 0$		0.5	$0.22\cot\alpha$	1.0	$0.44\cot\alpha$
Roller radial-thrust tapered		0.5	$0.22\cot\alpha$	1.0	$0.44\cot\alpha$

Table 6.82 Values of the coefficients X and Y for ball thrust-radial bearings. The values X, Y, and e for the contact angles α not mentioned in the table are determined from the given formulas. The ratio $F_a/F_r \le e$ is not used for single bearings. With $F_a/F_r > e$ it is assumed that $Y = 1$

α (°)	For single bearings with $F_a/F_r > e$	For double bearings with $F_a/F_r \le e$		$F_a/F_r > e$	e
	X	X	Y	X	
45	0.66	1.18	0.59	0.66	1.25
50	0.73	1.37	0.57	0.73	1.49
55	0.81	1.60	0.56	0.81	1.79
60	0.92	1.90	0.55	0.92	2.17
65	1.06	2.30	0.54	1.06	2.68
70	1.28	2.90	0.53	1.28	3.43
75	1.66	3.89	0.52	1.66	4.67
80	2.43	5.86	0.52	2.43	7.09
85	4.80	11.75	0.51	4.80	14.28
$\alpha \neq 90°$	$1.25 \tan\alpha \times (1 - 2\sin\alpha/3)$	$20 \tan\alpha/13 \times (1 - \sin\alpha/3)$	$10/13 \times (1 - \sin\alpha/3)$	$1.25 \tan\alpha \times (1 - 2\sin\alpha/3)$	$1.25 \tan\alpha$

the machines), noise level (for electric motors), constancy of the friction moment (for measuring apparatus and test equipment), or the value of the initial friction under load (for cranes), as well as on the actual geometry of the contact surfaces. The higher the listed requirements, the lower the value of the allowable static equivalent load.

If a high run smoothness is not needed, a short-term increase P_{0r} (P_{0a}) up to $2C_{0r}$ ($2C_{0a}$) is possible. With increased requirements of run smoothness, noise level, and constancy of the friction moment it is recommended that the allowable static equivalent load P_{0r} (P_{0a}) be reduced to C_{0r}/S_0(C_{0a}/S_0). The safety factor $S_0 = 1.5$ for thrust bearings of crane hooks and brackets, $S_0 = 2$ for precise instrumental equipment, and $S_0 = 4$ for important heavily loaded supports and turntables.

Specified Lifetime Calculation of Bearings

The basic data for this calculation are: F_{r1} and F_{r2}, the radial loads (radial reaction) of every support of the double-seat shaft (N); F_A, the external axial force acting on the shaft (N); n, the rotational frequency of the race (as a rule the rotational frequency of the shaft) (min^{-1}); d the diameter of the mounting shaft surface, which is taken from the layout diagram (mm); L'_{sa} and L'_{sah}, the required lifetime during which the probability of bearing operation failure is less than the appropriate probability, in millions of revolutions or hours, respectively; and the loading and operating conditions of the bearing unit (possible overload, working temperature, etc.).

Working conditions of the bearings are rather varied and can differ in terms of short-term overloads, working temperature, rotation of the inner or outer race, etc. The influence of these factors on the bearing efficiency is taken into account by means of the insertion of the equivalent dynamic load into the calculation.

As an equivalent dynamic radial (or axial P_a) load P_r one assumes a constant value that results in the same lifetime under the actual loading conditions.

The equivalent dynamic load is:

- *Radial*, for ball radial and ball or roller radial-thrust bearings

$$P_r = (VXF_r + YF_a)K_{dy}K_t \tag{6.29}$$

- *Radial*, for the roller radial bearings

$$P_r = F_r V K_{dy} K_t \tag{6.30}$$

- *Axial*, for ball and roller thrust bearings

$$P_a = F_a K_{dy} K_t \tag{6.31}$$

- *Axial*, for ball and roller thrust-radial bearings

$$P_a = (XF_r + YF_a)K_{dy}K_t \tag{6.32}$$

Here F_r and F_a are radial and axial loads on the bearing (N), X and Y are coefficients of the radial and axial dynamic loads, V is a coefficient of rotation ($V = 1$ for rotation of the inner race relative to the vector direction of the radial load, or $V = 1.2$ for rotation of the outer race), K_{dy} is a dynamic coefficient (Table 6.85); K_t is a temperature coefficient, its values are assumed depending on the operating temperature t_{oper} of the bearing: For operation under increased temperatures

Table 6.83 Values of the coefficient K_t

t_{oper} (°C)	K_t
≤ 100	1.0
125	1.05
150	1.10
175	1.15
200	1.25
225	1.35
250	1.4

bearings with a special stabilizing heat treatment or produced from heat-resistant steels are applied. The quality of the operation of the bearing under increased temperatures also depends on whether the lubricant used retains its properties, and on whether the materials of the seal and the retainer are chosen correctly.

The values X and Y depend on the class and structural features of the bearing, as well as on the ratio of the axial and radial loads. The limit value of the ratio F_a/F_r is a coefficient e of the axial loading.

For ball bearings with contact angle $\alpha < 18°$ the values of e are determined from the formulas given in Table 6.79 depending on the ratio $f_0 F_a/C_{0r}$. The values of the coefficient f_0 depending on the geometry of the bearing components and on the stress levels used in the calculation of the basic static radial load rating are given in Table 6.78 for ball radial and radial-thrust bearings.

The values of the coefficients X, Y, and e are assumed according to the data given in Table 6.79 for the

Table 6.84 Values of the coefficients X and Y for roller thrust-radial bearings ($\alpha \neq 90°$). The ratio $F_a/F_r \leq e$ is not used for the single bearings

Bearing classes	$F_a/F_r \leq e$		$F_a/F_r > e$		e
	X	Y	X	Y	
Single	–	–	$\tan\alpha$	1.0	$1.5\tan\alpha$
Double	$1.5\tan\alpha$	0.67	$\tan\alpha$	1.0	$1.5\tan\alpha$

Table 6.85 Recommended values of the dynamics factor K_{dy}

Load nature	K_{dy}	Application field
Quiet load without impulses	1.0	Low-power kinematic reduction gears and drives. Mechanisms of hand cranes, units. Power hoists, hand winches. Operating gears
Light impulses, short-time overloads up to 120% of the nominal load	1.0–1.2	Precise gearings. Cutting machines (except planing, slotting, and grinding machines). Gyroscopes. Lifting mechanisms of cranes. Telphers and monorail carriers. Winches with a mechanical drive. Electric motors with low and average power. Light fans and blowers
Moderate impulses, vibrational load, short-time overloads up to 150% of the nominal load	1.3–1.5	Gearings. Reduction gears of all types. Travel mechanisms of crane trolleys and swing-out mechanisms of cranes. Bushes of rail mobile trains. Boom changing mechanisms of cranes. Spindles of grinding machines. Electric spindles
Short-time overloads up to 180% of the nominal load	1.5–1.8	Centrifuges and separators. Boxes and propulsion engines of electric locomotives. Mechanisms and running wheels of cranes and road machines. Planers and slotting machines. Powerful electric machines
Loads with substantial impulses and vibrations; short-time overloads up to 200% of the nominal load	1.8–2.5	Gearings. Breaking machines and impact machines. Crank mechanisms. Rollers of rolling mills. Powerful fans
Load with strong impacts, short-time overloads up to 300% of the nominal load	2.5–3.0	Heavy forging machines. Log frames. Working roller conveyors of heavy section mills, blooming and slab mills. Refrigerating equipment

ball radial and radial-thrust bearings, in Table 6.80 for roller radial-thrust bearings ($\alpha \neq 0°$), in Table 6.82 for ball thrust-radial bearings, and in Table 6.84 for roller thrust-radial bearings ($\alpha \neq 90°$).

For bearings running under varying loading conditions, which are set with a load sequence diagram and with the loads corresponding to these rotational frequencies (Fig. 6.170), the equivalent dynamic load is calculated for the varying loading conditions using

$$P_E = \sqrt[3]{\frac{P_1^3 L_1 + P_2^3 L_2 + \ldots + P_n^3 L_n}{L_1 + L_2 + \ldots + L_n}} ,$$

where P_i and L_i are, respectively, the constant equivalent load (radial P_{ri} or axial P_{ai}) in the i-th mode of operation and the period of its action in millions of revolutions. If L_i is set in operating hours (L_{hi}), it is converted to millions of revolutions. Taking into account the rotational frequency n_i (min^{-1})

$$L_i = 60 n_i L_{hi} / 10^6 .$$

If the load on the bearing changes according to a linear law from $P_{\min}$ to $P_{\max}$, the equivalent dynamic load is

$$P_E = (P_{\min} + 2P_{\max})/3 .$$

It is known that the rates of machines with a varying load can be classified into six typical loading conditions (Sect. 6.2.8): 0 (constant), I (heavy), II (average equiprobable), III (average normal), IV (easy), and V (especially easy).

For bearings of shaft supports for toothed and worm gears running under typical loading conditions it is convenient to perform calculations with the help of the equivalence coefficient K_E:

Table 6.86 Values of the equivalence coefficient K_E

Mode of operation	K_E
0	1.0
I	0.8
II	0.63
III	0.56
IV	0.5
V	0.4

The equivalent loads are calculated by means of the known maximum long-term forces $F_{r1\,max}$, $F_{r2\,max}$, and $F_{A\,max}$ (corresponding to the maximum value of the long-term acting torques)

$$F_{r1} = K_E F_{r1\,max} , \quad F_{r2} = K_E F_{r2\,max} ,$$
$$F_A = K_E F_{A\,max} ,$$

according to which the bearing calculation is performed as though under constant load.

Matching of the rolling bearings is carried out in the following order:

1. The class and installation diagram of the bearing is set in advance.
2. For the appointed bearing the following data are written down from the manufacturer's catalog:
   – The values of the basic dynamic C_r and static C_{0r} radial load ratings for ball radial and radial-thrust bearings with contact angle $\alpha < 18°$; the main geometry: hole diameter d, outer diameter D, ball diameter D_w.

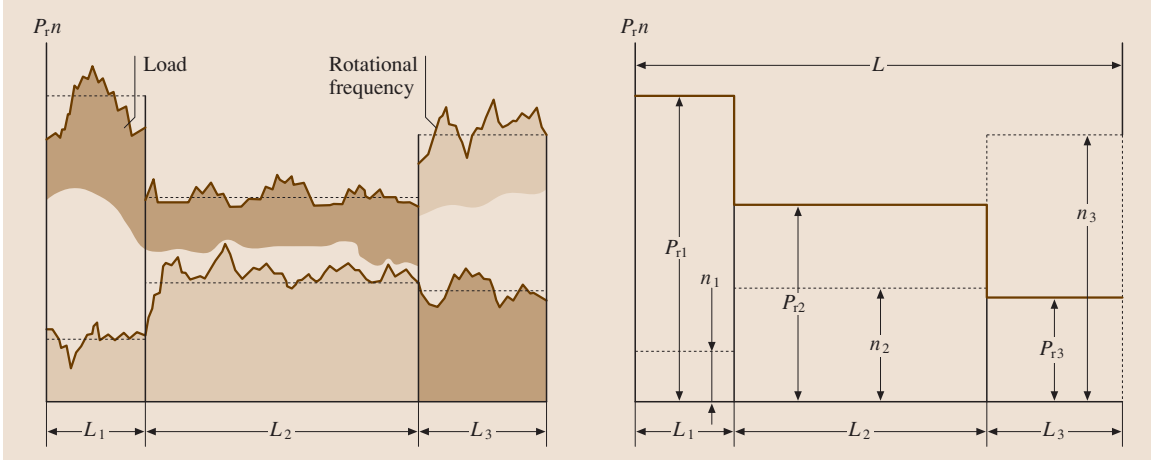


Fig. 6.170 Approximation of loads and rotational frequencies

- The value C_r for ball radial-thrust bearings with contact angle $\alpha \geq 18°$, and the values of the coefficients of the X radial, Y axial loads, the coefficient e of the axial loading from Table 6.79.
- The values C_r, Y, and e for tapered roller single-row bearings; $X = 0.4$ is also assumed (Table 6.80).

3. The axial forces F_{a1} and F_{a2} are determined from the equilibrium condition of the shaft and that of the minimum level of the axial loads on radial-thrust bearings.
4. For ball radial bearings, as well as for ball radial-thrust bearings with contact angle $\alpha < 18°$ the values X, Y, and e are determined according to Table 6.79, depending on the ratio $f_0 F_a / C_{0r}$. The values of the coefficient f_0 are given in Table 6.78 depending on the ratio $D_w \cos\alpha / D_{pw}$, where D_w is the ball diameter, α is the contact angle (for radial bearings $\alpha = 0°$), D_{pw} is the circle diameter of the center ball position: $D_{pw} = (d + D)/2$. In the absence of tabulated values the ball diameter can be calculated according to the height of *the effective cross-section* $H = (D - d)/2$:
   - For bearings from series 200, 300, and 400 for $d \leq 40$ mm for the *especially easy* series $D_w = 0.6H$.
   - For bearings from series 200, 300, and 400 for $d > 40$ mm $D_w = 0.635H$.
   - For compact and high-speed bearings $D_w = 0.55H$.
   - For bearings of increased load rating $D_w = 0.64H$.
5. The ratio F_a/F_r is compared with the coefficient e, and the values of the coefficients X and Y are finally assumed: for $F_a/F_r \leq e$ it is assumed that $X = 1$ and $Y = 0$, for $F_a/F_r > e$ for ball radial and radial-thrust, and roller bearings the earlier (under points 2 and 4) values of the coefficients X and Y are finally assumed.
6. The equivalent dynamic load is calculated ((6.29)–(6.32)).
7. The design lifetime of the bearing, which has been corrected according to the safety level and use conditions, is determined (h)

$$L_{sah} = a_1 a_{23} \left(\frac{C}{P}\right)^k \frac{10^6}{60n},$$

where C is a basic dynamic load rating of the bearing (radial C_r or axial C_a) (N), P is an equivalent dynamic load (radial P_r or axial P_a, and under varying loading conditions P_{Er} or P_{Ea}) (N), k is an exponent that takes on the value $k = 3$ for ball bearings and $k = 10/3$ for roller bearings, n is the rotational frequency of the race (min^{-1}), a_1 is the coefficient adjusting the lifetime depending on the required safety (Table 6.41), and a_{23} is a coefficient adjusting the lifetime depending on special properties of the bearing, which it obtains, e.g., as a consequence of the application of special materials or special production processes or special structure, as well as its working conditions (conformity of the lubricant characteristics with the required ones, the presence of the foreign particles causing behavioral changes of the material).

The basic design lifetime is confirmed based on the test results of the bearings on special machines and in certain conditions characterized by the presence of a hydrodynamic oil film between the contact surfaces of the races and the solids of revolution and by the absence of increased warp of the bearing races. Under real operating conditions deviations from these conditions are possible, which are approximately estimated by using the coefficient a_{23} (Table 6.40).

With the choice of the coefficient a_{23} the following use conditions of the bearing are distinguished:

a) Common (material of usual fusion, presence of the race warps, absence of a safe hydrodynamic oil film, and presence of foreign particles)
b) The presence of the elastic hydrodynamic oil film in the contact between the races and the solids of revolution, the absence of increased warps in the unit; standard production steel.
c) The same as in item (b), but the races and the solids of revolution are manufactured from steel of electroslag or vacuum-arc refining.

Design formulas for lifetime are correct for rotational frequencies over $10\,\text{min}^{-1}$ to the limit frequencies according to the manufacturer's catalog, and also if P_r (or P_a), and with varying loads $P_{r\,max}$ (or $P_{a\,max}$) does not exceed $0.5C_r$ (or $0.5C_a$).

In some cases, the allowable load P_r (or P_a) is determined from the formula for the lifetime calculation. For bearings running with low rotational frequencies and those intended for a short lifetime the allowable load calculated in such a way can exceed the static load rating, which is inadmissible. Thus, adaptability of the formulas is restricted by the condition $P_r \leq C_{0r}$ (or $P_a \leq C_{0a}$).

8. The fitness of the planned dimension type of the bearing is estimated. The bearing is suitable if the design lifetime L_{sah} is more than or equal to the

Table 6.87 Recommended values of the design lives of machines and equipment

Machines, equipment, and their operating conditions	Lifetime (h)
Devices and equipment used occasionally (demonstration equipment, domestic appliances, devices, technical plants for medicine purposes)	300–3000
Mechanisms used during a short period of time (agricultural machines, lifting cranes in assembly workshops, light conveyors, construction machines and mechanisms, electric hand tools)	3000–8000
Important mechanisms running with breaks (auxiliaries in power stations, conveyors for flow-line production, lifts, not often used metal-working machines)	8000–12 000
Machines for one-shift operation with underload (fixed electric motors, reduction gears of general industrial function, rotor crushing plants)	10 000–25 000
Machines running under full load during one shift (working machines, woodworkers, machines for general engineering, lifting cranes, separators, centrifuges, fans, conveyors, graphic arts equipment)	20 000–30 000
Machines for round-the-clock use (gear-drives of roller mills, compressors, mine hoists, fixed electric machines, ship drive, pumps, textile equipment)	40 000–50 000
Wind power plants, including the main shaft, gearboxes, generator drives	30 000–100 000
Hydroelectric power plant, rotating furnaces, machines for high-speed cable winding, motors for ocean liners	60 000–100 000
Continuously running machines with high load (equipment for paper-making plants, electric power plants, mine pumps, equipment of merchant ships, rotary furnaces)	≈ 100 000

required one

$$L_{sah} \geq L'_{sah} \, .$$

In some cases, two identical radial or radial-thrust single-row bearings are installed together in one support. If the bearings are manufactured precisely and assembled so that they run as a unit, this pair is considered as one double-row bearing. For the lifetime determination from the formula of item (7) the basic dynamic radial load rating $C_{r\,sum}$ of the set of two bearings is substituted for C_r, taking the value $C_{r\,sum} = 1.625C_r$ for ball bearings and $C_{r\,sum} = 1.714C_r$ for roller bearings. The basic static radial load rating of this set is equal to twice the nominal load rating of a single-row bearing $C_{0r\,sum} = 2C_{0r}$.

For the determination of the equivalent load P_r the values of the coefficients X and Y are assumed as for double-row bearings: for ball bearings according to Table 6.79; for roller bearings according to Table 6.80.

If the bearing unit comprises two self-contained bearings, which are substituted independently of each other, these premises are not applicable.

The recommended values of the bearing lifetime of different machines and equipment are given in Table 6.87.

6.11.15 Fits of Bearing Races

Bearing races can be classified into the following categories: local, circulating, and oscillatory.

Local loading applies when when the resulting radial load acting on the bearing is always supported by the same limited section of the rolling path of the race and is transmitted to a corresponding part of the mounting surface of the shaft or the casing.

Circulating loading applies when the resulting radial load acting on the bearing is supported and transmitted through the solids of revolution to the rolling path in a rotational process in sequence along its whole length and, therefore, along the whole mounting surface of the shaft or the case.

Oscillatory loading applies when the fixed race of a bearing is subjected to the influence of the resulting radial load, which therefore performs periodic oscillatory motion.

For circulating loading the connection of the races with the shaft or the case should be made through interference, which prevents turning and running of the mated component with the race and consequently beading of the mounting surfaces, contact corrosion, galling, decrease of rotational accuracy, and imbalance.

Under local loading, fits that allow small clearance are used. Running of the components mated with the races does not occur under such loading, and casual turning of the nonrotary race is useful, as the position of its loading zone changes. Furthermore, this mating facilitates axial displacements of the races in assembly, clearance adjustment in the bearings, and thermal deformations.

Races under oscillatory loading and with clearance of the mated component also exhibit relative motion, as in the case of circulating loading. Displacement occurs only over a certain surface area. This also causes contact corrosion and wear, which is why races subjected to oscillatory loading are installed using interference.

Bearing fits can differ from standard fits in terms of the position and the values of the tolerance ranges for the mating surfaces of the races. The required fits in the connections of rolling bearings are obtained by setting the appropriate tolerance ranges for the shaft diameters and the holes in the case. Figure 6.171 shows the position of the most commonly used tolerance ranges relative to the bearing hole and its outer surface. One peculiarity consists of the fact that the tolerance range for the hole diameter in rolling bearings is not positioned from the zero line (not *on the positive side*), but down (*on the negative side*). In this way interferences are guaranteed in the connections of the inner race with the shafts, with the tolerance ranges k, m, and n. The tolerance range for the diameter of the outer race is positioned as usual *on the negative side* or *in the component body*.

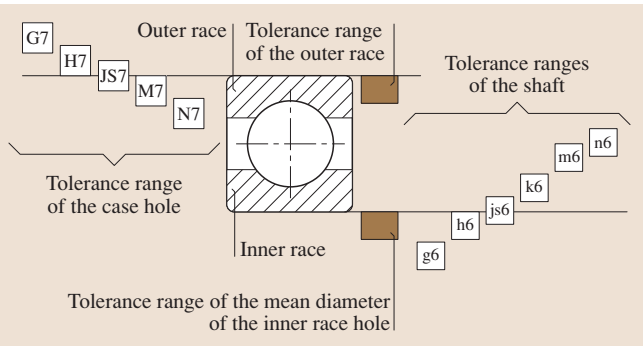


Fig. 6.171 Position of the tolerance ranges relative to the hole of the bearing and its outer surface

The value of the equivalent dynamic load P also influences the choice of the fit. In accordance with this we have the following loading conditions

$$\begin{aligned} &\text{Light :} && P \le 0.07C \\ &\text{Normal :} && 0.07C < P \le 0.15C \\ &\text{Heavy :} && P > 0.15C \end{aligned}$$

where C is the dynamic load rating of the bearing. The heavier the conditions, the tighter the fit. As a rule, roller bearings run under high loads, so the fits of roller bearings are tighter than those of ball bearings. In Russia the fits of the races on the shaft and in the case are determined according to the recommendations of [6.89]. The fit is often chosen by an analog method to obtain analogous long-working checked components similar in structure, function, and operating conditions.

6.12 Design of Bearing Units

6.12.1 Clearances and Preloads in Bearings and Adjustment of Bearings

General radial or axial displacement of one race of bearings relative to another is considered a radial or axial clearance. The optimal values of the radial and axial clearances for given operating conditions of the bearing allow for efficient load distribution between the solids of revolution, and between the required displacement of the shaft and the case in the radial and axial directions, as well as improve and increase the vibroacoustic stability and reduce friction losses.

In nonadjustable bearings classes there are three kinds of radial clearances: initial, setting, and operating. The setting clearance is always less than the initial one because of race deformations in the radial direction after installation of the bearing at the work site. During operation of the bearing unit and under steady temperature conditions the operating clearance forms, which can be more or less than the set value depending on the installation configuration of the bearings, taking into account the load and temperature differences applied to the shaft and case.

In Russia, in accordance with [6.114], values of the radial and axial clearances for a rolling bearing as delivered, and reference designations of the clearance groups *normal and additional* (with lower or higher clearance values) are specified.

Table 6.88 Values of radial clearances for ball radial single-row bearings without grooves for insertion of balls with a cylindrical hole according to [6.114]

Nominal hole diameter d (mm)	Clearance G_r (µm)									
	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max
	Clearance group									
	6		Normal		7		8		9	
Over 10 to 18	0	9	3	18	11	25	18	33	25	45
Over 18 to 24	0	10	5	20	13	28	20	36	28	8
Over 24 to 30	1	11	5	20	13	28	23	41	30	53
Over 30 to 40	1	11	6	20	15	33	28	46	40	64
Over 40 to 50	1	11	6	23	18	36	30	51	45	73
Over 50 to 65	1	15	8	28	23	43	38	61	55	90
Over 65 to 80	1	15	10	30	25	51	46	71	65	105
Over 80 to 100	1	18	12	36	30	58	53	84	75	120
Over 100 to 120	2	20	15	41	36	66	61	97	90	140

Nonadjustable bearing classes are produced with comparatively small clearances after installation on the shaft when they can run without additional adjustment.

Table 6.88 provides the dimensions of the radial clearances for ball radial single-row bearings with a cylindrical hole as an example.

Bearings intended for regular operating conditions (temperature difference between the outer and inner races of 5–10 °C, temperature of the inner race usually higher than the temperature of the outer one) should have a clearance corresponding to the basic, normal group.

The application field of bearings with increased clearances is supports with substantial operating temperature fluctuations, as well as supports where the races of the bearing are assembled on the shaft, and the case with considerable fit interference, because of the expected high dynamic loads, or supports where warps of the inner races relative to the outer races are possible as a consequence of manufacturing errors, assembly, or insufficient shaft rigidity. Radial single-row ball bearings being loaded only with axial forces should also have increased radial clearance, which allows the contact angle in the bearing to be increased, i. e., its axial load rating increased. Radial non-self-installed bearings with increased radial clearance are also used under deviation from slots coaxiality.

Bearings with a cut clearance are installed in supports with high requirements for the radial or axial runout of shafts operating with moderate rotational frequency with efficient cooling, as well as in supports where greater heating of the outer races than of the inner ones is expected as a consequence of particular heat sources.

The axial and radial clearances of adjustable bearings can be fixed in certain ranges only by assembly in the machine unit. The required axial clearance in thrust bearings is settled the same way as for assembly.

The optimal value of the clearances is fixed experimentally for each particular unit. If the bearings are assembled with a large clearance, then the whole load is taken by one or two balls or rollers (Fig. 6.172a). The working conditions of bearings with large clearances are not favorable, which is why these clearances are not admissible. A decrease in clearances results in a more even load distribution between the solids of revolution, reduces vibrations, and increases rigidity of the support. The presence of some axial clearances positively influences the reduction of the antitorque moment. Standard radial-thrust bearings are adjusted in such a way that the axial clearance is about zero under steady tem-

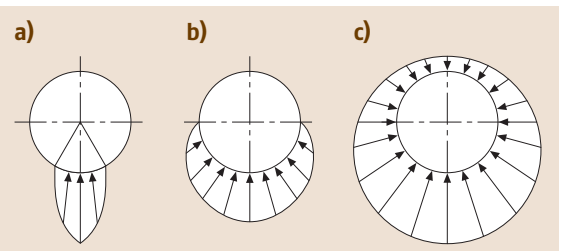


Fig. 6.172a–c Load distribution on the solids of revolution (a) for increased clearance, (b) for zero clearance, (c) for preload or substantial axial load

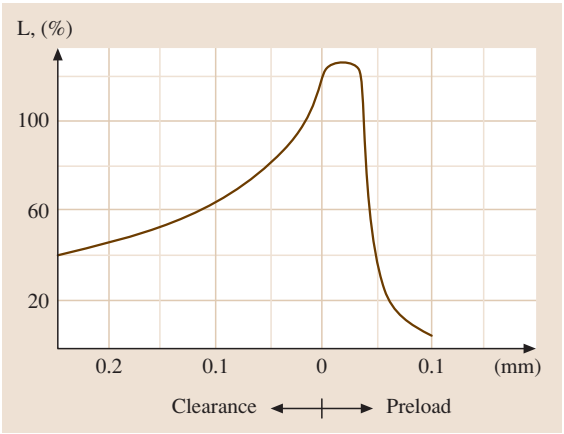


Fig. 6.173 Influence of clearances and interferences on the lifetime (L is a lifetime in percentage of the rated value)

perature conditions. In this case, about one-half of the solids of revolution are under the action of the radial load (Fig. 6.172b).

In some units, e.g., in the machine tool industry, assembly of the bearings with a preload is used to increase support rigidity and the rotational accuracy of the shaft, and to improve the vibroacoustic feature of the unit. In this case, more than one-half or all the solids of revolution of the bearing are under loading (Fig. 6.172c).

The nature of the preload consists in the fact that a pair of bearings has previously been loaded by the axial force, which removes the axial clearance in the set making an initial resiliency at the contact points of the working surfaces of the races with the solids of revolution. If the working axial force is then applied to the bearing, the relative displacement of its races will be lower, as a consequence of the extra deformation of the working surfaces, than before the preload. The preload causes an identical deformation in both bearings. With too great an interference the bearings run under hard usage, as the loads on the solids of revolution, antitorque moment, and wear increase, and the bearing lifetime is therefore reduced.

The influence of the clearance interferences on the lifetime L is shown in Fig. 6.173.

Ball radial, radial-thrust, and roller tapered bearings are installed with a preload. This is also so with bearings with short cylindrical rollers mountable on a conic shaft journal with interference, which is able to cause an extension of the inner race and completely remove radial clearance in the bearing.

The recommended axial clearances for single-row radial-thrust ball and roller tapered bearings are given in Tables 6.89 and 6.90, respectively. The values in these tables correspond to the standard operating conditions, when the temperature of the inner races of the bearings does not exceed that of the outer races by more than 10 °C, and the differential temperature of the shaft and the case amounts to ≈ 10–20 °C.

By assembling the bearing with a flare for a conical shaft journal the initial radial clearance decreases as a consequence of the extension of the inner race. The axial displacement of the inner race with a hole having a taper of 1:12 relative to the shaft journal or the clamping sleeve causes a reduction of the initial radial clearance, being approximately 1/15 of the displacement value.

Adjusting Methods

Adjustment is the installation of the inner clearance in the bearing or installation of the preload in the bearing unit. The radial and axial clearances in the bearing are mutually dependent. With a clearance change in one direction (e.g., in the axial direction) the clearance also changes in the other (radial) direction. The clearances in the bearings are made and changed on assembly of

Table 6.89 Recommended axial clearances (μm) for ball radial-thrust single-row bearings. Installation configurations of the bearings: 1 – two in a support; 2 – one in every support

Hole diameter of the bearing d (mm)		Axial clearance by contact angle α (°)					
		12				26 and 36	
		Configuration 1		Configuration 2		Configuration 1	
Over	To	Min	Max	Min	Max	Min	Max
–	30	20	40	30	50	10	20
30	50	30	50	40	70	15	30
50	80	40	70	50	100	20	40
80	120	50	100	60	150	30	50
120	180	80	150	100	200	40	70
180	260	120	200	150	250	50	100

Table 6.90 Recommended axial clearances (μm) for radial-thrust roller tapered single-row bearings. Installation configurations of the bearings: 1 – two in a support; 2 – one in every support

Hole diameter of the bearing d (mm)		Axial clearance by contact angle α (°)					
		10–16				25–29	
		Configuration 1		Configuration 2		Configuration 1	
Over	To	Min	Max	Min	Max	Min	Max
–	30	20	40	40	70	–	–
30	50	40	70	50	100	20	40
50	80	50	100	80	150	30	50
80	120	80	150	120	200	40	70
120	180	120	200	200	300	50	100
180	260	160	250	250	350	80	150
260	360	200	300	–	–	–	–
360	400	250	350	–	–	–	–

the product most often by means of axial displacement of the outer and inner races or (rarely) by means of radial deformation of the inner race by its fit onto the cylindrical or bevel surface of the shaft.

A radial preload is usually used in roller bearings with cylindrical rollers, double-row radial-thrust ball bearings, and sometimes in radial ball bearings. For example, the preload is applied with the help of the interference fit of a sufficient size of one or two races of the bearing, where the initial radial inside clearance in the bearing decreases to zero. As a result in operation the clearance becomes negative, i. e., a preload appears. Bearings with a flare are the most convenient for applying a radial preload, as the force of the preload can be adjusted rather exactly by moving the bearing along its bevel mounting surface (on the shaft journal, clamping sleeve, or tightening bushing).

The axial force of the preload required for single-row radial-thrust ball bearings, tapered roller bearings, and radial ball bearings is made by means of the displacement of one of the races relative to the other along the axis by a distance corresponding to the required force of the preload.

Two fundamentally different main adjusting methods are applied: individual and combined adjustment.

With individual adjustment each bearing unit is regulated separately with the help of nuts, washers, spacing, deformable sleeves, etc.; changing and checking allow the nominal value of the preload force to be maintained with the lowest possible deviations. The following measuring procedures of the preload are used:

- According to displacement, which is determined by means of the component measurement of the bearing unit, taking into account the thermal expansion of the components in operation and a certain force loss of the preload during some operation time, i. e., taking into account the resiliency in the system.
- According to the frictional moment with the use of the known ratio between the bearing load and frictional moment in it. This method is universal, requires little time, and can be easily automatized.
- According to the directly measured force, which can be made or changed by adjustment.

In practice, the first two methods are used more often due to their simplicity and availability.

For combined adjustment all of the components of the bearing unit must be completely interchangeable, which in the end results in a tightening of their dimensional tolerances.

The advantage of individual adjustment is that single unit components can be manufactured according to free tolerances (e.g., corresponding to the 9th–14th accuracy degree) and the preload is applied with a comparatively high degree of accuracy.

6.12.2 Principal Recommendations Concerning Design, Assembly, and Diagnostics of Bearing Units

Design Recommendations

The design of a product should be adapted for convenient assembly, and precise installation and dismantling of the bearing units.

The mounting surfaces of shafts and cases should have hollow chamfers or contact lead-ins with a small taper angle to guarantee precise prior centering, decrease shearing and bearing microasperity, and a smooth insertion force increase with assembly.

The hollow chamfer radii of the shafts and the cases at the mounting sites must provide a reliable fit of the race face to the supporting shoulder surface of the shaft or the case. It is also necessary that perpendicularity of the supporting shoulder surface of the shaft or the case to the common axis be ensured.

The race of the bearing cannot be installed with interference, but should be applied without a thrust block into the mated component, otherwise the runout of the face appearing in the pressing process may substantially exceed the allowable values. Shoulders or spacer rings of insufficient height cannot be used. In Russia the shoulder dimensions must conform to the [6.113].

The dimensions of the elements of the shaft and the case mated with the bearings must guarantee reliable support of the bearing races when subjected to axial forces and the possibility of using a press or mechanical removers that grab the extension of the race above the shaft shoulder or the case hole to dismantle them must be available.

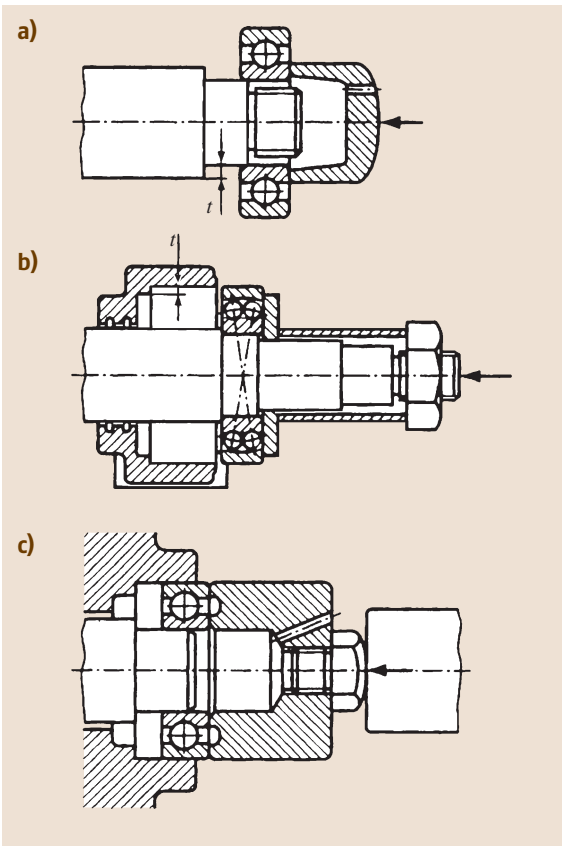


Fig. 6.174a–c Installation of the bearings **(a)** onto the shaft, **(b)** into the case, **(c)** simultaneously onto the shaft and into the case

Assembly and Dismantling of Bearings

Upon installation (or dismantling) of the bearings on the shaft and into the case the condition that the axial force must be applied directly to the race that is being built up (or dismounted) must be met. The force must not be transmitted through the solids of revolution (balls and rollers) upon assembly and dismantling of the bearing. Otherwise, hollows may appear on the rolling paths and solids of revolution. It is impossible to apply the assembly forces to the cage.

Figure 6.174 shows some installation methods of the bearings on the shaft (Fig. 6.174a), into the case (Fig. 6.174b), and simultaneously on the shaft and the case (Fig. 6.174c). The bearing races have low rigidity. For correct installation the race should be inserted up to the stop in the shoulder. The height t of the shoulders on the shafts and in the holes of the cases or the sleeves (Figs. 6.174 and 6.175) defines the dimension r of the race bevel. The shoulder height must form a sufficient supporting surface for the faces of the bearing races. The minimum height t of the shoulders is assumed as:

Table 6.91 Recommended minimum height t of the shoulders

r (mm)	t (mm)
0.5	1.0
1.0	1.8
1.5	2.5
2.0	3.0
2.5	4.0
3.0	4.8
3.5	5.5
4.0	6.5

The shoulder height is usually assumed to be equal to half of the race width. The holes in the assembly sleeves (Fig. 6.174a,c) are intended for the outlet of air from the inner hollow of the sleeve by the bearing insertion on the shaft.

For the assembly of bearings, mechanical erection tools (spanners, impact, socket wrenches), and hydraulic (hydraulic nuts, thrusts) or heating (inductive, electric ovens with thermostat, oil reservoir) units can be used as well as presses.

Figure 6.176 shows an installation diagram for a bearing with a conical hole on the shaft using hy-

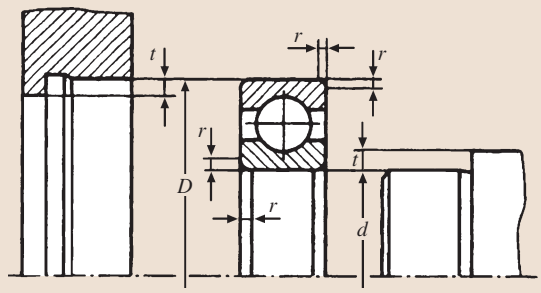


Fig. 6.175 Determination of the collar height

draulic thrust. Oil under pressure is supplied with a plunger pump through the hole into the shaft groove under the inner race, which is burst opened. By means of rotation of the nut the bearing is moved in the axial direction to the installation site. Bearings with a cylindrical hole are mounted in the same way. However, for assembly of a bearing on a cylindrical area they are necessarily inserted up to the stop in the shaft shoulder.

It is evident from this figure that, on assembly of bearing with hydraulic thrust, the following must be considered in the shaft structure: a threaded section for the nut, a threaded hole for the connecting pipe of the oil-duct (M6 by $d \leq 100$ mm), a hole ⌀2.5 mm, and a groove (width 3 mm, depth 0.5 mm) for the oil supply.

For assembly of bearings of the open *class* with a cylindrical hole on the shaft with interference it is advisable that the bearing be heated first. The required differential temperature Δt between the bearing race and the shaft or the case depends on the interference value of the fit and on the bore diameter of the race. The

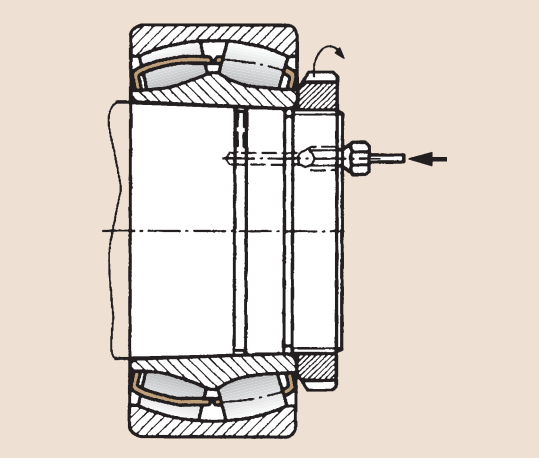


Fig. 6.176 Installation of the bearing using hydraulic thrust

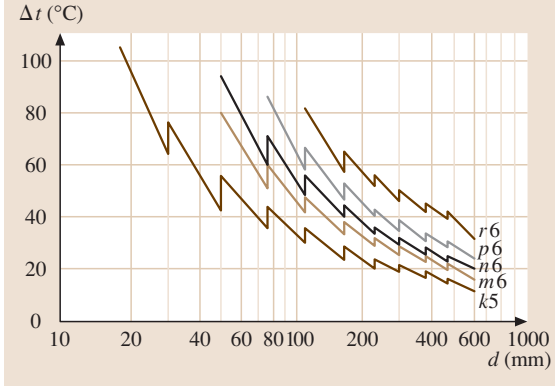


Fig. 6.177 Approximate values of the differential temperature Δt between the bearing race and the shaft for the most commonly used tolerance ranges depending on the inner diameter d

approximate values of the differential temperature Δt for the most widely used tolerance ranges depending on the hole diameter d can be determined from Fig. 6.177. It is inadmissible to heat the bearings to a temperature above 125 °C, as otherwise changes of the material structure of the bearing may occur.

Uniform heating can be achieved with the help of electric heat boosters, heating furnaces, and oil reservoirs. In the latter case, the bearing is dipped into the reservoir with clean mineral oil, which has a high flash point, heated, e.g., up to 80–90 °C and held for 10–15 min. For assembly of bearings with safety washers and permanently laid lubricant heating is carried out in a thermostat.

For bearing installation into the case with interference it is recommended that the bearing be cooled previously to a temperature of −70 to −75 °C in a thermostat by using dry ice or to heat the case to 20–50 °C higher than the temperature of the bearing.

For dismantling of small ($d < 80$ mm) and average ($d = 80$–200 mm) bearings *spiral removers* are used: with two (Fig. 6.178a) or three folding rods (Fig. 6.178b,c) (the maximum dismantling force is 6–50 kN, and with a hydraulic booster up to 80 kN).

The remover in (Fig. 6.178c) also allows the use of two rods for dismantling, which are mounted in two large bosses. The installation sites of the bearings must be physically designed in such a way to that the removers can be operated with convenience.

When being removed from the case, the bearing must be grabbed by the outer race (Fig. 6.179a, and by dismounting off the shaft, by the inner race Fig. 6.179b).

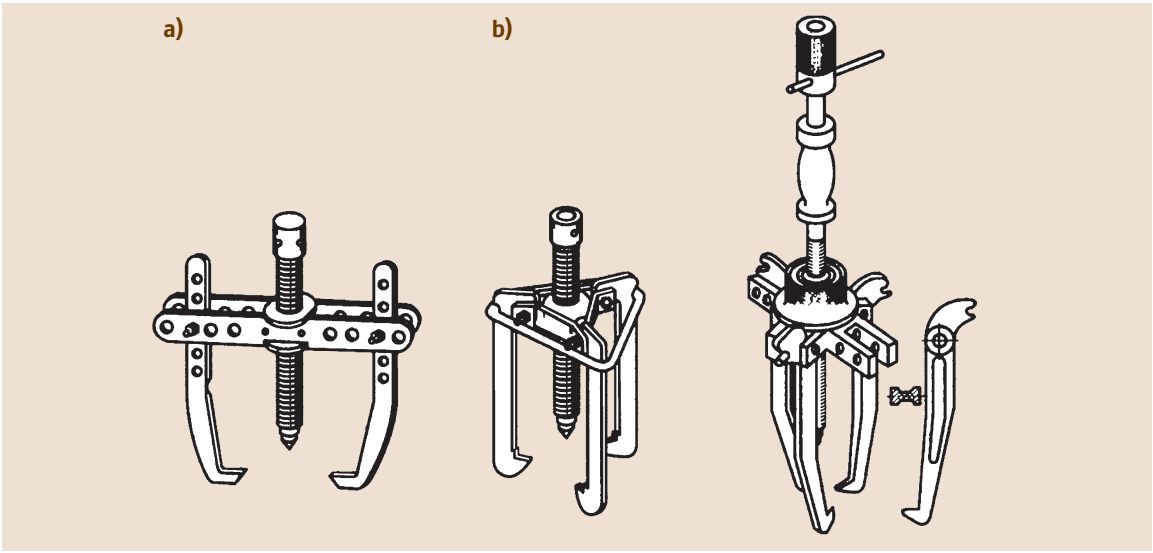


Fig. 6.178a,b Screw removers with (a) two and (b) three rods

In order for the bearing race to be grasped by the rods, the height of the remover t of the shoulder (Fig. 6.179a) must not be too high. The minimum dimension of the t_1 inner face and t_2 outer face extension of the bearing race intended for dismantling is as follows:

Table 6.92 Recommended minimum dimensions, t_1 – inner face and t_2 – outer face

Shaft diameter (mm)	$t_1 = t_2$ (mm)
< 15	1
15–50	2
> 50–100	3.5

With high shoulders it is necessary to foresee grooves for the position of the remover rods (Fig. 6.179b – a remote element A).

For location of the remover rods (Fig. 6.179a) a free space $a \approx (0.4–0.5)C$ is included, where C is the width of the bearing race on removal of the outer race from the blind hole.

Diagnostics of the rolling bearing state during operation and planning of their maintenance is gaining increasing importance. If damage of the bearings can be determined in the initial stage, they can be replaced in time by carrying out a scheduled repair of the machine, and thus extraordinary machine lockup can be prevented.

For estimation of the current status of the running bearing units without lockup and dismantling of the product, and for timely detection of the first features of beginning fracture (flaking on the working surfaces of the races and the solids of revolution) various systems and devices are used; most of them are based on the measurement and analysis of vibrations.

However, in practice not all machines or machine units are controlled by means of modern devices. The machine operator must pay attention to the indirect features of possible damage, e.g., increased noise, temperature, or vibration of the bearing units.

Noise control. The most common method of assessing the status of bearings is listening. With the help of an ascultoscope (most commonly, by means of a wooden stick) increased noise can be detected, and the experienced operator can determine its source. In a faultless state the bearings generate a quiet buzzing noise. Clattering, hissing, or any unusual noise indicates improper status of the bearings.

Temperature control. The integrated indicator of the quality and work stability of the bearing unit is its temperature. Increased temperature indicates abnormal operation of the bearing and, moreover, has a negative influence on the lubricant characteristics and its intake into the bearing. Continuous running at a temperature over 125 °C can result in a decrease in the lifetime of the bearing.

The reason for the higher temperature or its constant change can be insufficient lubrication and excess lubrication, increased loads, lubricant contamination, too little clearance in the bearing, excessive interference, in-

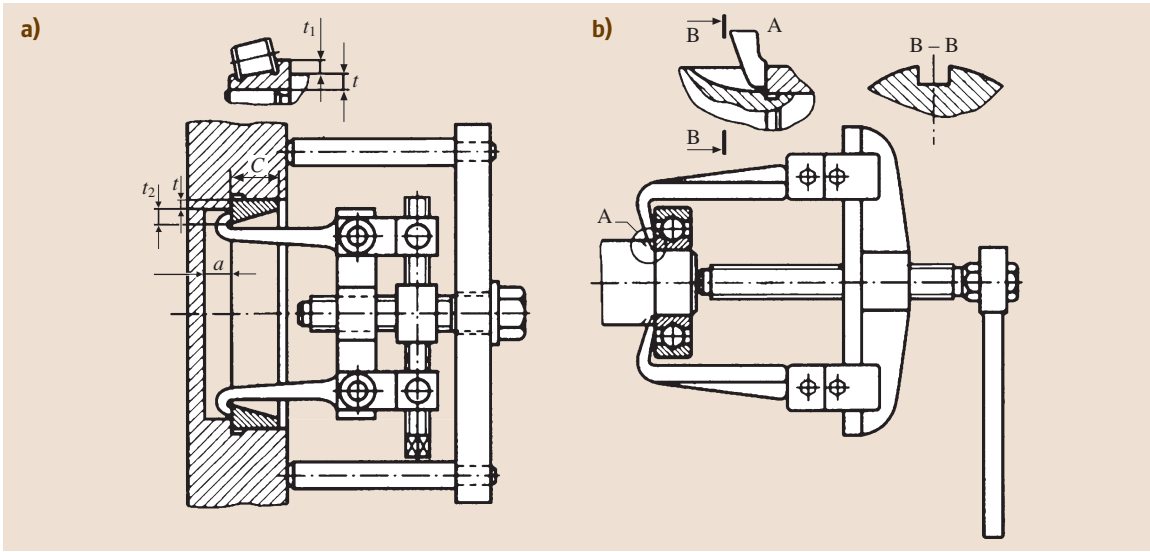


Fig. 6.179a,b Dismantling of the bearing from (a) the case and (b) the shaft

creased frictional moment as a consequence of higher mutual race warp, low-quality manufacture of the mating components, or heightened friction in the seals.

6.12.3 Design of Bearing Units

Figure 6.167 shows the main diagrams for axial shaft fixation. It is more convenient to consider the designs of the bearing units for each diagram separately for the fixing support and the floating support.

Fixing Support in Diagram 1a (Fig. 6.167)

With the axial fixing of the shafts according to diagram 1a, the bearing classes shown in Fig. 6.180 are used in the fixing supports.

Fastening of the Bearings on the Shafts

Figure 6.181 shows methods of fastening bearings on the shaft, which are applied with shaft loading with a substantial axial force in both directions.

Reliable fastening of the bearing is carried out with a round slotted nut (Fig. 6.181a), which is stopped from spontaneous breakout with a multitab washer. The retainer has one inner jut and six outer jut tabs. The inner jut of the washer gets to a specially made groove on the shaft, and one of its outer juts is turned back into the slot of the washer.

Fastening with an end plate is simple and reliable (Fig. 6.181b). In this case, the pin fixes the washer from turning relative to the shaft. So that the end plates do not cause unbalance under high rotational frequencies, they

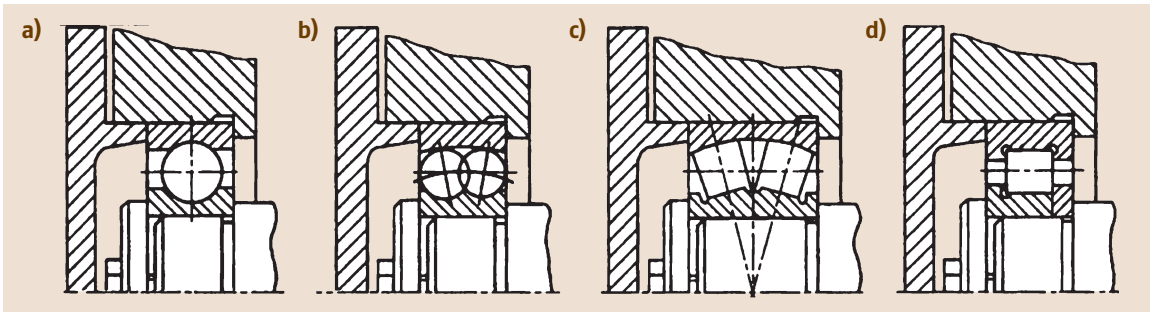


Fig. 6.180a–d Design of the fixing support in diagram 1a (Fig. 6.167) using the following types of bearings: (a) ball radial single-row, (b) ball radial spherical double-row, (c) roller radial spherical double-row, (d) with short cylindrical rollers and an extra race

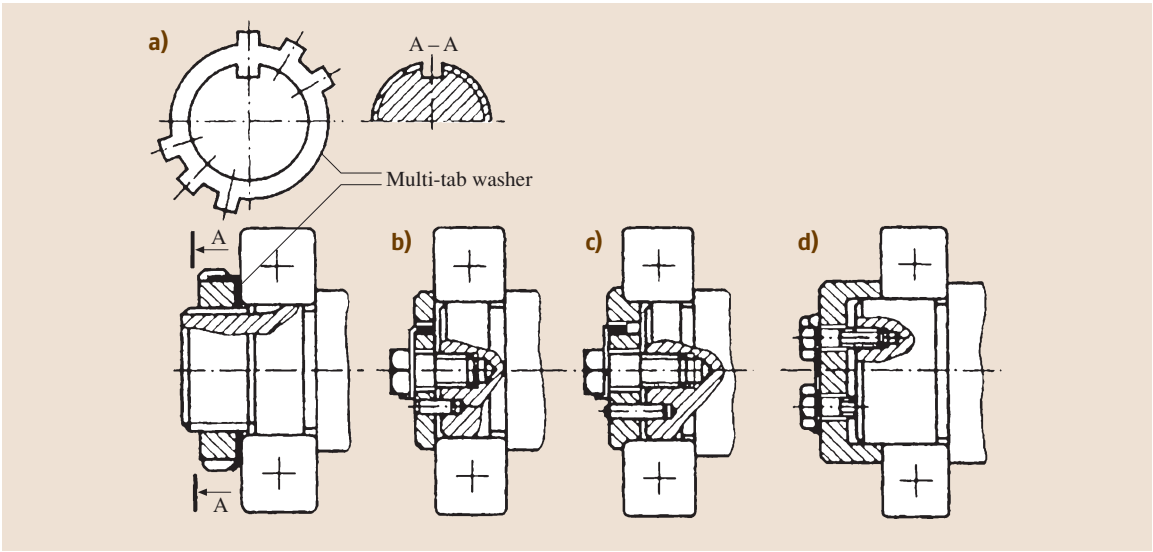


Fig. 6.181a–d Fastening means of the bearings on the shaft. **(a)** With a round slotted nut, **(b)** with an end plate, **(c)** with an end plate centered according to the hole of the bearing, and **(d)** with an end plate centered according to the shaft

are centered according to the bearing hole (Fig. 6.181c) or to the shaft (Fig. 6.181d). In all variants it is necessary to consider stopping of the screws, which fix the washer to the shaft face, from self-unfastening. In Fig. 6.181b,c stopping of the screw is carried out by means of a retainer with a toe, and in Fig. 6.181d by means of a deformable washer, installed under both screws at once. The ends of the washer are turned back on the panes of the screw heads.

Fastening of the bearings with a spring thrust planar outer ring is finding increasing application (Fig. 6.182c).

The races of the rolling bearings are manufactured in width b (Fig. 6.182a) with rather great deviations. Thus with a hole diameter of over 30–50 mm the width tolerance is 0.12 mm, and with a diameter of over 50–80 mm it is 0.15 mm. The shaft dimension e performs with approximately the same accuracy. The thickness of the spring ring s has a tolerance of 0.12 mm. The clearance z between the thrust ring and the bearing is $z = e - s - b$.

The presence of the clearance z, which can vary in range from 0 to 0.3 mm for bearings with a hole diameter, e.g., over 50–80 mm, is a disadvantage of the given fastening. To avoid this effect, it is advisable to install a compensatory ring (2) between the bearing and the spring thrust ring (1) (Fig. 6.182b). The clearance is minimized by means of matching of this ring according to the thickness or extra machining in compliance with measuring results during assembly.

The ends of special tongs, with which the rings are opened, are placed into the holes of the spring rings (Fig. 6.182b) on their removal from the shaft. The thickness of the spring rings is small, so the tongs do not

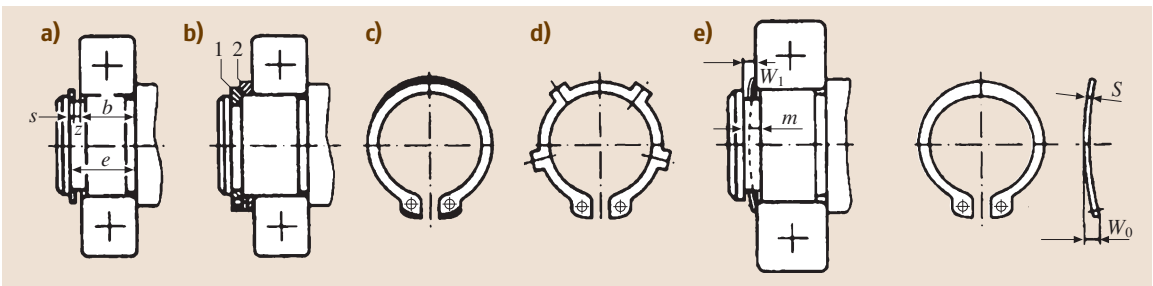


Fig. 6.182a–e Fastening of the bearing with a spring ring. **(a–c)** Planar, **(d)** tab, **(e)** bent

penetrate the hole very deeply and often come off. In order to avoid this, the groove is milled on the ring face (2) (Fig. 6.182b).

It should be borne in mind that the spring thrust ring projects lightly above the shaft surface. Thus, the eccentric ring overlaps a chamfer of the bearing only on a small surface (shaded area in Fig. 6.182c). On the greater part of the circle the spring planar ring does not touch the race face of the bearing at all, which is why the compensatory ring not only decreases the axial clearance, but also improves the contact of the bearing with the spring ring.

The firm SEEGER (Germany) and others use tab spring rings (Fig. 6.182d), which make contact with the race of the bearing at six points. The same firms apply bent spring thrust bearings for tightening of the bearings to the face of the shaft shoulder (Fig. 6.182e) that exclude application of other compensators. Compensative abilities of such rings are characterized by the following data (mm):

Table 6.93 Main parameters of the bent spring rings

d (mm)	40 – 100
s (mm)	1.75 – 3.0
m (mm)	3.4 – 6.3
W_0 (mm)	3.5 – 6.9
W_1 (mm)	$2.1^{+1.2} - 3.3^{+2.4}$

It should be taken into account that spring thrust planar rings can transmit considerable axial forces. Thus, for example, the allowable axial force for the spring thrust planar ring amounts to 17.1 kN with a shaft diameter of 30 mm.

Making of the Thrust Collars on the Shaft

A structural peculiarity of the rolling bearing is the fact that its inner race is a rather compliant component. The inner race has to be tightened by assembly to the shaft collar or the face of the component, which is mounted on the shaft, to be installed precisely on the shaft without warp. The race of the bearing must fit to the thrust collar with its flat face. On the one hand, the height of the shaft collar must be higher than the coordinate of the bearing chamfer, while on the other hand, it must be chosen taking into account the possibility of removal of the bearing from the shaft. The required information concerning the height of the shaft collar is given above.

If, for some reason, a shaft collar of the required height cannot be made, then one of the following variants is used:

- The spacing washer of the required height is mounted between the shaft collar and the bearing race (Fig. 6.183a).
- The collar is made by means of installation of the spring thrust planar ring into the shaft groove (Fig. 6.183b).
- The extra ring (1) is mounted, which improves contact of the bearing with the spring ring (Fig. 6.183c).
- Two half-washers of hooked type or of rectangular cross-section are installed in the groove on the shaft, which the inner race of the bearing (Fig. 6.183d),

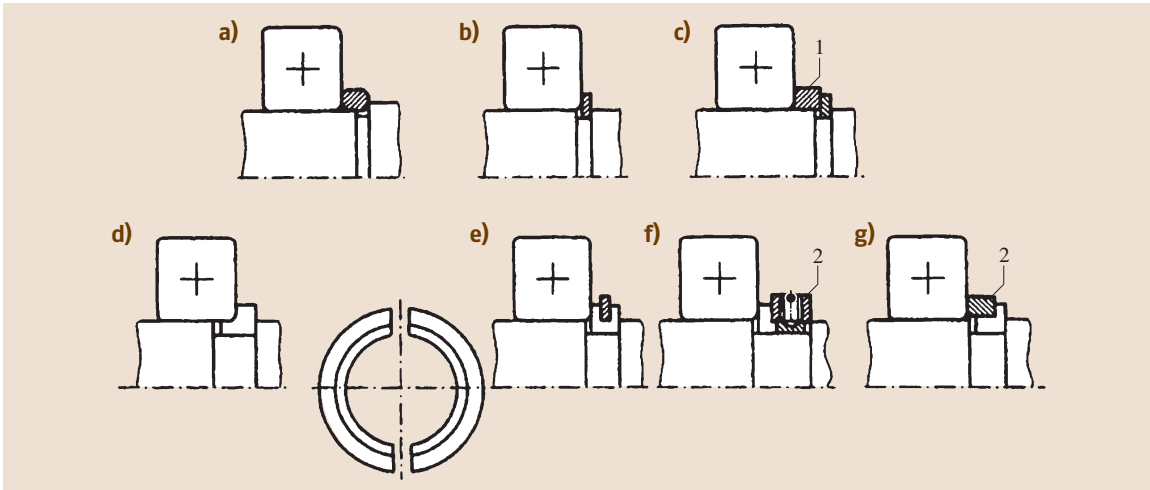


Fig. 6.183a–g Methods of making the collars on the shaft. **(a)** With a spacing washer, **(b–d)** with a planar spring ring, **(c)** with two semirings of L-type section, **(d)** with two semirings secured with a spring ring, **(f,g)** with one-piece ring

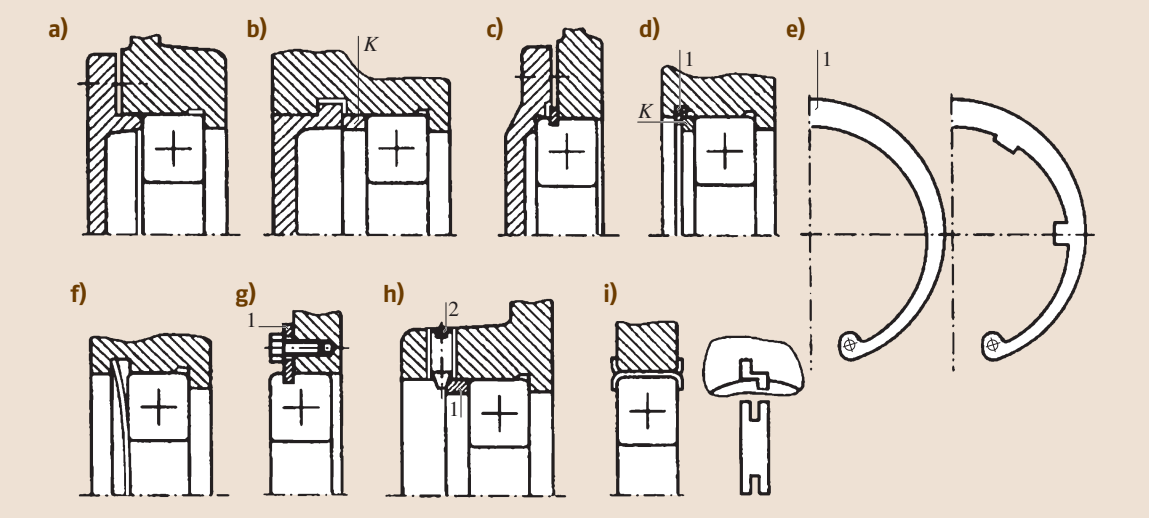


Fig. 6.184a–i Fastening of the bearings in the case. **(a)** With a clamp-on cap, **(b)** an insert cap, **(c)** with a thrust washer on the outer race of the bearing, **(d)** with a spring ring: planar, **(e)** tab, **(f)** bent, **(g)** with two semirings, **(h)** with three set screws, **(i)** with two plates with adjustable ends

the spring ring (Fig. 6.183e), and the nondetachable ring 2 (Fig. 6.183f,g) prevent from falling out.

Fastening of the Bearings in the Case

Figure 6.184 shows the most frequently used methods for fastening the bearings in the case. A simple and reliable method of fastening the bearing in the case *with a cap* is widely used, either clamped (Fig. 6.184a) or inserted (Fig. 6.184b). It is simpler to fix bearings that have a groove for the thrust ring on the outer race of the bearing. The spring thrust planar ring (Fig. 6.184c) or two half-washers (Fig. 6.184g) are installed into the groove and fastened on the case with screws. The advantage of these methods is the fact that the hole of the case does not have a ledge to complicate its machining.

In (Fig. 6.184d) the bearing is fixed with a spring thrust planar ring (1). To fasten the race of the bearing in the case without clearance the compensatory ring K is sometimes installed between the thrust ring and the bearing. To improve contact with the bearing, race tab spring rings are used (Fig. 6.184e). A compensator is not needed to fasten the bearing with a spring bent retaining ring (Fig. 6.184f), which tightens the outer bearing race to the case collar.

Figure 6.184h shows fastening of the bearing with the help of three set screws and a ring (1). To apply this method it is necessary to have the possibility to position three screws evenly along the case circle. The lock ring (2) prevents the screws from self-unfastening.

In lightly loaded supports fastening is done with the help of plates with adjustable ends in the absence of axial forces (Fig. 6.184i). Two plates are usually installed at 180° along the circle. The plates are placed into the axial grooves on the mounting hole of the case. The ends of the plates are bent in pairs onto the case and the outer race of the bearing.

All of these fastening means for the bearing in the case are more or less equivalent.

Making of the Thrust Collar in the Case

For precise installation the outer races of the bearings are tightened to the collar of the housing part. According to Fig. 6.185a the thrust collar is made directly in the case. However, the presence of a ledge in the hole of the housing part presents certain problems with hole boring. Hole machining of the housing part can be simplified by making a collar in the sleeve (Fig. 6.185b). However, incorporation of an additional laborious and precise component – a sleeve – can be avowed only in the case that the sleeve helps to solve another design problem, e.g., simplification of assembly or making of the free-standing assembly unit.

Execution of the collar by means of erection of a spring thrust planar inner ring (Fig. 6.185c) is easier. It should be taken into account that the spring rings can transmit substantial axial forces. Thus, e.g., the allowable axial force for the spring thrust planar ring becomes 74.7 kN with a hole diameter of 62 mm.

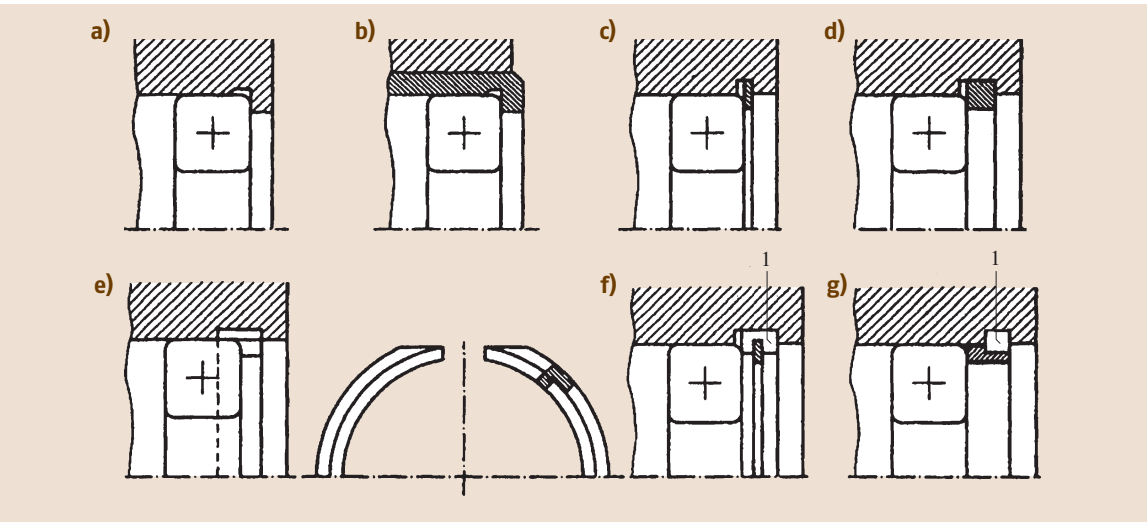


Fig. 6.185a–g Methods of making the collars in the case. **(a)** With a shoulder in the case, **(b)** with a shoulder in the sleeve, **(c)** with a planar spring ring, **(d)** with an unbroken ring in a split frame, **(e)** with two semirings of L-type section, **(f)** two semirings secured with a spring ring, **(g)** with a one-piece L-type ring

In the case with a split along the shaft axis the thrust collar can be made with *an unbroken ring*, which is put into the groove of the case hole (Fig. 6.185d).

In Fig. 6.185e the thrust collar is made with *two semirings* of hooked-type cross section. Semirings are placed into the groove of the case hole. The chamfers on the semirings enable their installation into the hole groove of a one-piece case. The outer race of the bearing prevents the semirings from falling out. The collar shown in Fig. 6.185f is made with two semirings, which are prevented from falling out of the case groove by a spring thrust ring. Two semirings forming a thrust collar in the variant illustrated in Fig. 6.185g are kept from falling out by the one-piece ring of hooked-type cross section. All the thrust collars made according to Fig. 6.185 are able to support substantial axial forces and can be used with any of the fastening means of the bearings shown in Fig. 6.184.

Adjustment of Axial Clearances in Bearings

When a shaft fixing in a support with one bearing (diagram 1a in Fig. 6.167) adjustment is not carried out. The axial clearance is made when the bearing is manufactured.

A fixing support is shown in diagram 1b (Fig. 6.167). When the axial shaft is clamped as shown in diagram 1b, the bearing classes are applied in the fixing supports, which are given in Fig. 6.186a–i. The thrust collars on the shafts and in the holes of the housing parts are designed according to one of the variants shown in Figs. 6.183–6.185.

The angular rigidity of the fixing supports, where the bearings are positioned in compliance with the variants in Fig. 6.186b,d,f,h, is higher than that of supports with positioning of the bearings according to the variants in Fig. 6.186a–c,e,g.

In some bearing classes (e.g., radial, radial-thrust ball, radial spherical ball, and roller bearings) the axial clearances between the rings and the solids of revolution are made by production of the bearings. In others (tapered roller bearings) the axial clearances are set on assembly.

Clearance adjustment of radial or radial-thrust bearings of the fixing support in diagram 1b (Fig. 6.167) is carried out by means of the axial displacement of the outer and inner races.

Adjustment of Bearings by Means of Axial Displacement of the Outer Races

Figure 6.187a shows the adjustment with a gasket package (1), which is installed under the flange of the bearing cap. A package of thin (thickness ≈ 0.1 mm) metallic gaskets is applied to it. It is also convenient to adjust with a gasket package of different thicknesses. The Timken firm (USA) delivers the following gasket package: three pieces with thickness 0.127 mm, three pieces with thickness 0.179 mm, or one piece with thickness 0.508 mm. Rather precise adjustment

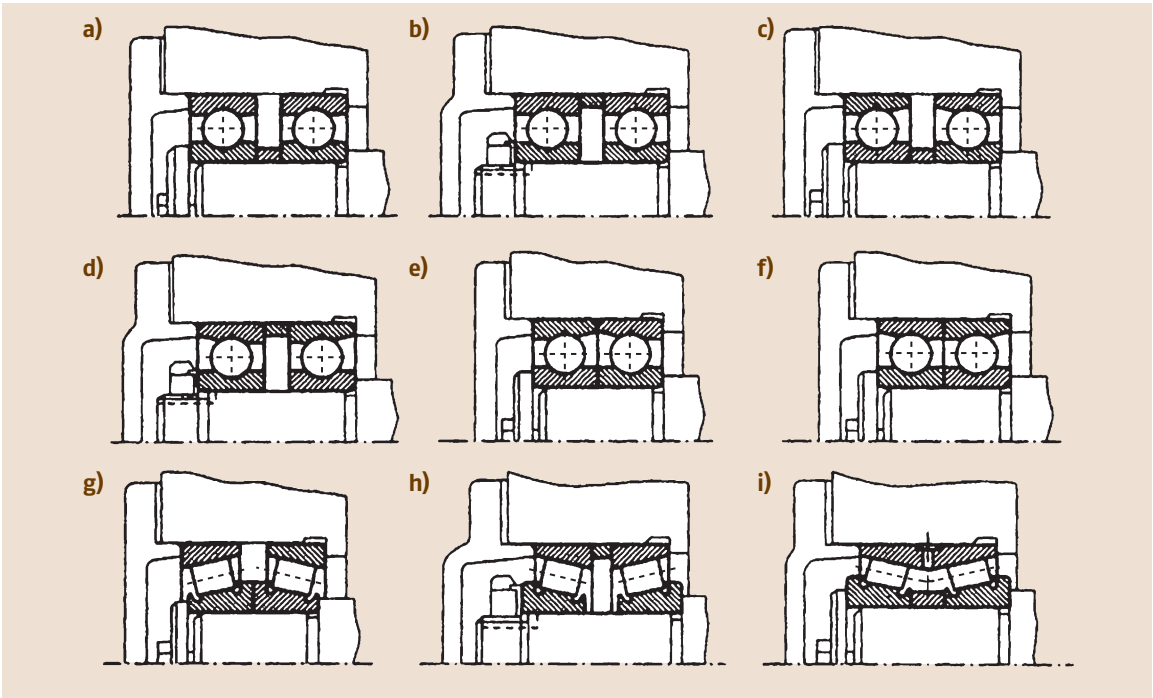


Fig. 6.186a–i Design of the fixing support in diagram 1b (Fig. 6.167) using the following types of bearings: **(a,b)** ball radial, **(c–f)** ball radial-thrust, **(g,h)** roller tapered, **(i)** roller-tapered double-row

can be obtained with a gasket package with the thickness series: 0.05, 0.1, 0.2, 0.4, and 0.8 mm. Calculation of the required adjustment with the gasket package is performed by means of a probability-theoretical method. Sometimes, instead of using the gasket package, adjustment is carried out with two semirings that are installed under the flange without removal of the cap.

Adjustment of the bearings can be carried out with a stud screwed into the case (Fig. 6.187b). It should be taken into account that, in this case, the positioning accuracy of the bearing is decreased. The positioning accuracy can be increased with the help of the influence of the screw (1) on the washer (2) (Fig. 6.187c). The washer self-installs along the face of the outer bearing race due to the presence of the spherical surface on the screw face (1). In this design the washer (2) should be stiff, and the diameter of the adjusting screw should be as large as possible. For smaller diameters of the screws there have been cases of screw extraction from the bearing cap under the action of axial forces. Adjustment accuracy in the configuration shown

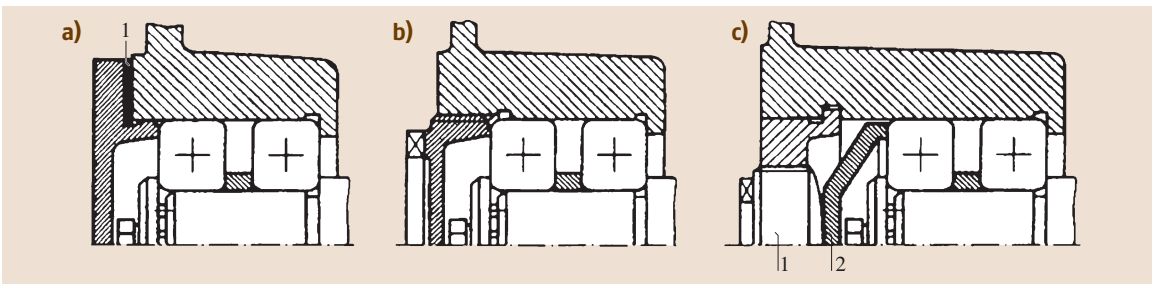


Fig. 6.187a–c Adjustment methods of bearings with axial displacement of the outer races. **(a)** With metallic gaskets, **(b)** with a stud screwed into the case, **(c)** with a screw and a self-installed washer

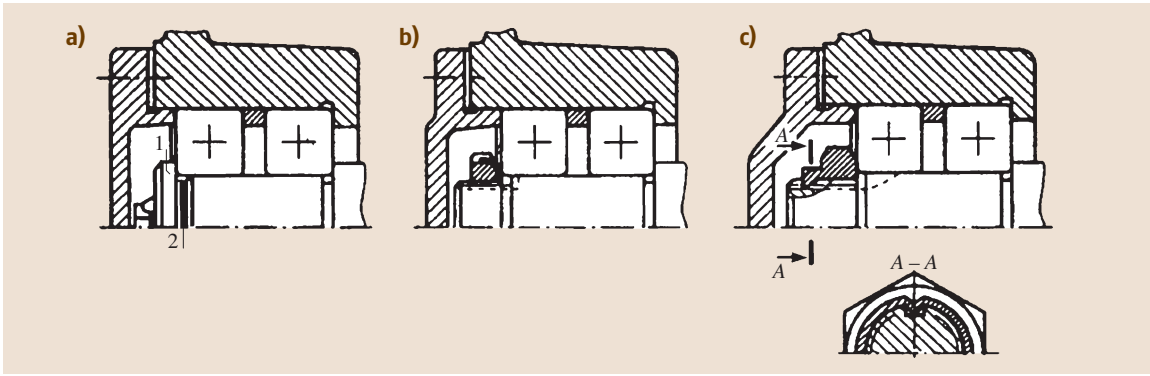


Fig. 6.188a–c Adjustment methods of bearings with axial displacement of the inner races. **(a)** With metallic gaskets, **(b)** with a round slotted nut, **(c)** with a nut with deformable shoulder

in Fig. 6.187c can be increased by decreasing the thread pitch. Thus, in these structures threads with a fine pitch are used.

Adjustment of Bearings with Axial Displacement of the Inner Races

In Fig. 6.188a adjustment of the bearings is carried out by means of tightening the face washer (1). A package (2) of thin metallic gaskets is installed between the faces of the shaft and the washer. The washer is fixed to the shaft face with a screw and locked.

In Fig. 6.188b the bearings are adjusted with a nut. After making the required clearance in the bearings the slotted nut is locked with a multitab washer. To this end the nut must be installed in such a way that the slot on it coincides in position with one of the bending ledge tabs of the retainer. In some cases, this condition results in degradation of accuracy of the adjustment. Adjustment with a nut with a special circular deformable shoulder (Fig. 6.188c) does not suffer from such a disadvantage. Two grooves are made (at 180°) on the thread shaft. After making the required clearance in the bearings the nut is locked, pressing the edges of the deformable shoulder into the shaft grooves.

Practice indicates that it is not necessary to loosen the fit under a moving inner race for adjustment. Adjustment of the bearing is an important operation. The quality of the adjustment depends on the professional skills of the assembler. Bearings can easily be undertightened or overtightened. Thus, on some plants the required rigidity is obtained by means of matching and grinding of the spacing washers (1 and 2), which are installed between the bearings on the shaft and in the case (Fig. 6.189). Then both the inner and outer races of the bearings are fixed on the shaft and in the case. This method is very reliable, but requires precise measuring of the bearing dimensions and thorough fitting of the races.

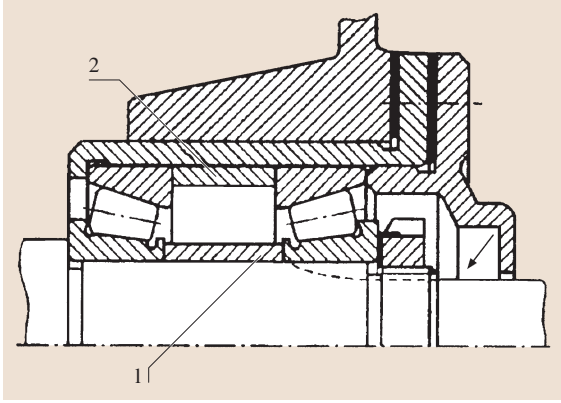


Fig. 6.189 Adjustment of bearings with spacing washers

Floating Supports in Diagrams 1a and 1b (Fig. 6.167)

For axial fixation of the shaft according to diagrams 1a and 1b the bearing classes given in Fig. 6.190a–h are used in the floating supports. The clearance b is included between the faces of the outer bearing race and the cap in the floating support. The value of the clearance in the supports designed according to Fig. 6.190a–c can be assumed to be $b \geq 0.01l$, where l is the distance between the faces of the bearing races (mm) (Fig. 6.167). In a support designed according to Fig. 6.190e it is assumed that $b \approx 0.5–0.8$ mm. To fasten the races on the shafts or the housing parts the methods given in Figs. 6.181–6.185 can be used.

Fig. 6.190a–h Design of the floating supports in diagrams 1a and 1b using the following bearings: **(a)** ball radial single-row, **(b)** ball radial spherical double-row, **(c)** roller radial spherical double-row, **(d,e)** with short cylindrical rollers, **(f,h)** needle ▶

Adjustment of the Bearings

The rigidity of the floating support can be increased by means of special design methods. Figure 6.191a,b shows floating supports where permanent interference is provided with the installation of the race (1), with a large number of springs positioned along the circle.

The firm SKF (Sweden) recommends that interference in the bearings be made with both coil and disk springs (Fig. 6.191a–c). In the latter case, the springs guarantee a constant force. The required radial stiffness of the floating support in the bridge mill (Fig. 6.191d) is obtained by straining of the inner race on the bevel neck.

Supports with a Preload

Preloading of the bearings is usually carried out by means of mutual axial race displacement (shown schematically in Fig. 6.192). The forming diagram of the preload is similar in the case of the installation of gaskets, springs, or races of unequal thicknesses. The preload is used to increase the stiffness of both fixing and floating supports.

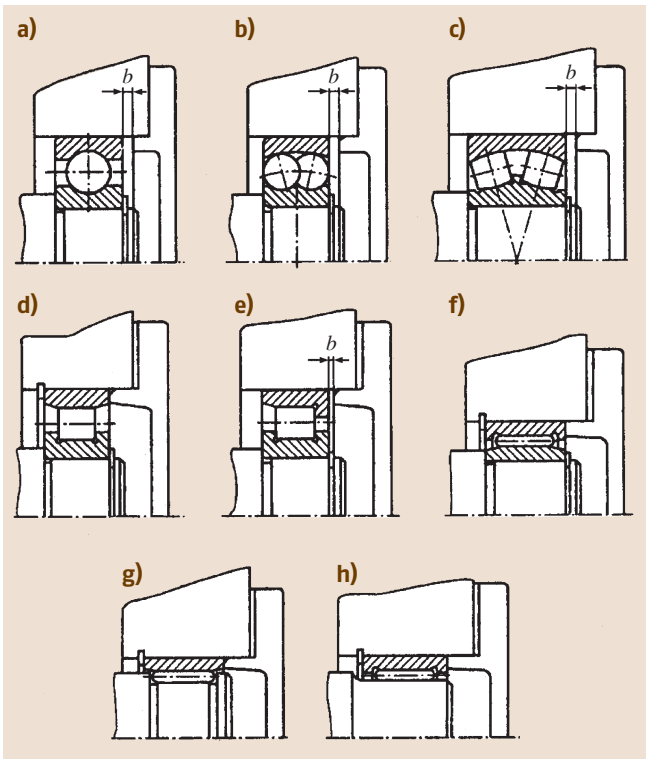
Preload of Bearings of Fixing Supports

Figure 6.193 shows the principal methods of preload making in the bearings of the fixing supports in diagram 1b. The preload is made by means of face grinding of the inner races (Fig. 6.193a) by the value required for obtaining the set interference after axial compression of the outer and inner races with each other; with the help of gaskets (Fig. 6.193b) or races of different thicknesses (Fig. 6.193c); as well as springs (Fig. 6.193d).

The preloading of the bearings of the floating supports can be applied with compression rings (Fig. 6.194a), the application of rings with different thickness (Fig. 6.194b), grinding of the faces of the inner races (Fig. 6.194c), and by a specially matched ring (1) (Fig. 6.194d).

Supports According to Diagram 2a (Fig. 6.167)

With the axial shaft fixation according to diagram 2a both supports are designed equally. Figure 6.195a–h shows examples of the embodiment of one shaft support; other supports are designed in a similar manner.



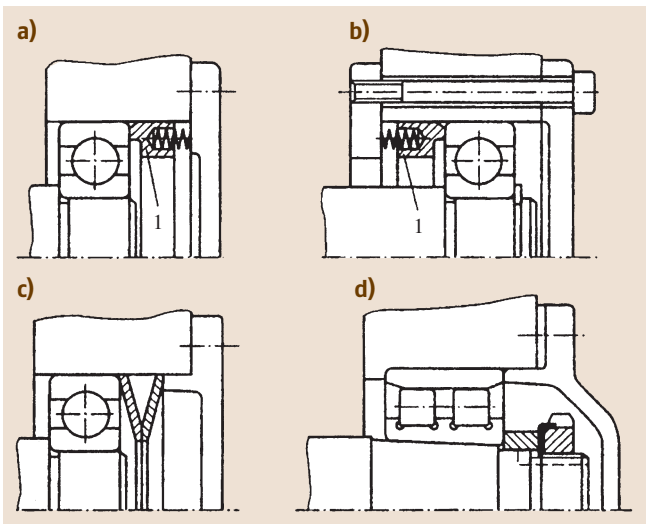


Fig. 6.191a–d Methods of increasing the stiffness of the floating support. **(a,b)** With a ring with a large number of coil compression springs, **(c)** disk springs, **(d)** with deformation on the cone of the inner bearing race

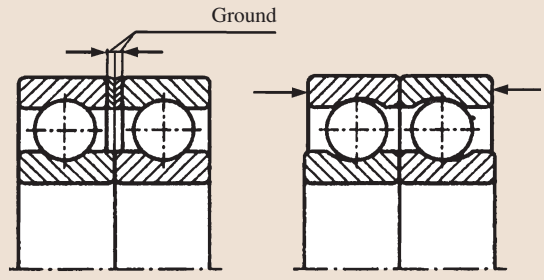


Fig. 6.192 Forming diagram of the preload

Clearance adjustment in the bearings is carried out by means of axial displacement of the outer races. In Fig. 6.196 adjustment with the help of a package of thin metallic gaskets (1) is shown, which are installed under the flanges of the caps of the clamped-on bearings. To adjust the bearings the package of gaskets can be installed under the flange of one of the caps. If it is further necessary to adjust the axial shaft position, the entire package of gaskets is halved, and then each of them is mounted under the flange of the corresponding cap. Adjustment with a package of metallic gaskets guarantees a rather high accuracy; this method is used for the installation of both radial and radial-thrust bearings.

Through the application of inserted caps, adjustment of the radial bearings can be carried out by means of the installation of the compensatory ring (1) between the faces of the outer race and the inserted cap (Fig. 6.197a). For convenience of assembly the compensatory ring is mounted from the side of the blind bearing cap. Upon mounting of the radial ball bearings a clearance $a = 0.2–0.5$ mm is left between the face of the outer race and the face of the bearing cap to compensate for thermal deformations (Figs. 6.196 and 6.197a). This clearance is not shown in the drawings of the assembly units in view of its insignificance.

Adjustment of radial-thrust bearings in the case of the application of inserted caps is carried out according to Fig. 6.197b, operating the screw (1) on the self-installed washer (2). To increase the adjustment accuracy threads with a fine pitch are used.

When the operating mode of the product is modified, the temperature changes and, therefore, so does the clearance in the bearings and their rigidity. After a time, the adjustment of the bearings carried out during assembly gradually disintegrates due to wear and crumpling of microridges. Thus, periodic readjustment of the bearings is needed.

More or less permanent support rigidity is obtained with the help of the application of resilient members

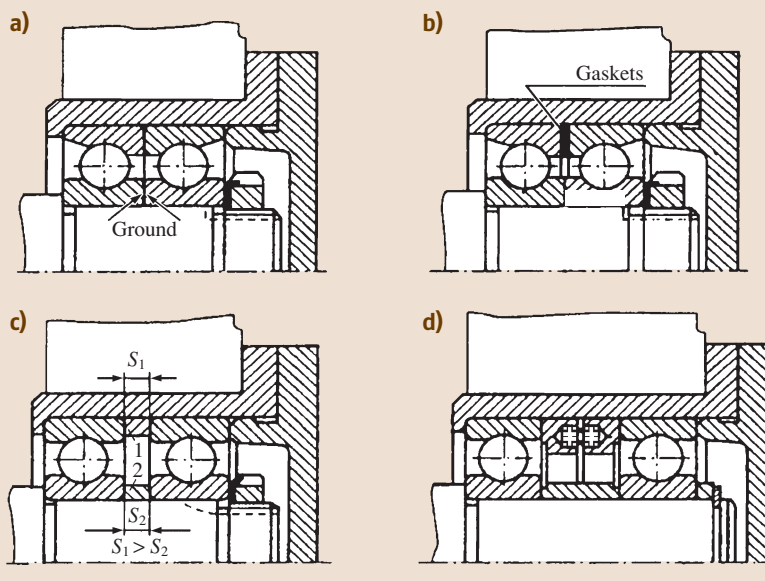


Fig. 6.193a–d Forming of the preloads in the fixing supports. **(a)** With grinding of the faces of the inner races, **(b)** mounting of metallic gaskets between the outer races, **(c)** with rings of different thickness, **(d)** rings with compression springs

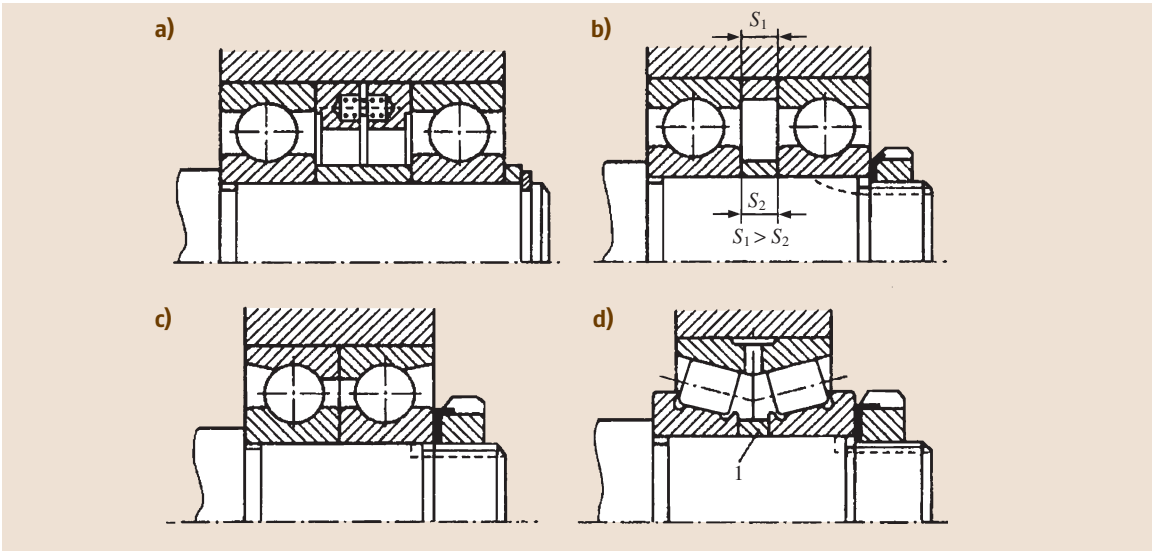


Fig. 6.194a–d Forming of the preloads in floating supports. **(a)** With rings with compression springs, **(b)** with rings of different thickness, **(c)** with grinding of the faces of the inner races, **(d)** with a specially matched ring (1)

(Fig. 6.198), which compensates for wear. The springs are positioned along the circle and mounted in the ring (1) (Fig. 6.198a,b). In the bearings of the firm Game (France), as well as in Russian bearings, the outer race is joined with a ring (1) (Fig. 6.198c). The width of the outer race is increased, which increases the positioning accuracy of the bearing along the hole surface of the housing part.

The resilient members are incorporated into the support on which the axial force does not have any in-

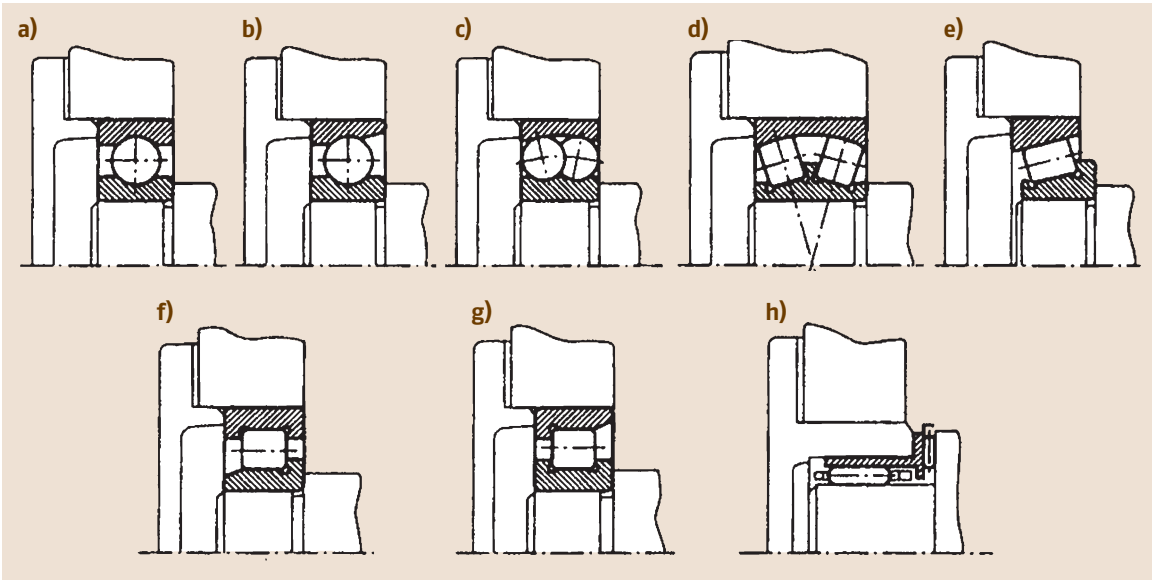


Fig. 6.195a–h Methods of executing the fixing support in diagram 2a using the following bearings: **(a)** ball radial single-row, **(b)** ball radial-thrust, **(c)** ball radial double-row spherical, **(d)** roller radial double-row spherical; **(e)** tapered roller; **(f,g)** with short cylindrical rollers, **(h)** needle combined

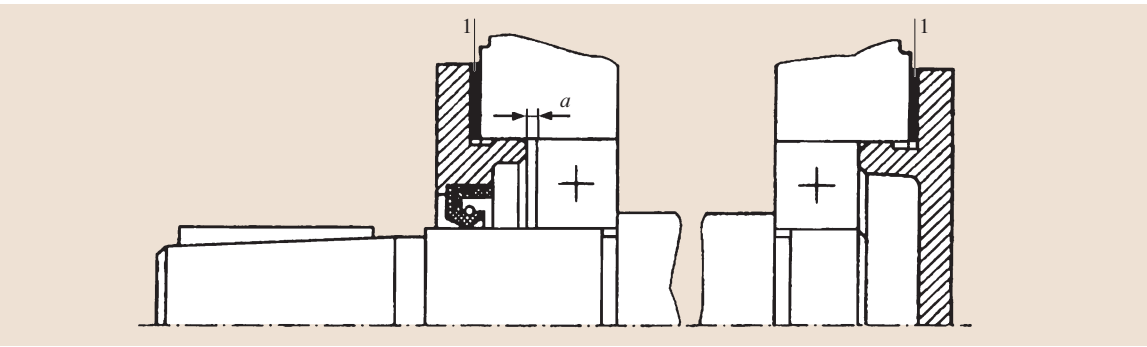


Fig. 6.196 Adjustment of clearances in the bearings (diagram 2a) with a set of thin metallic gaskets (1)

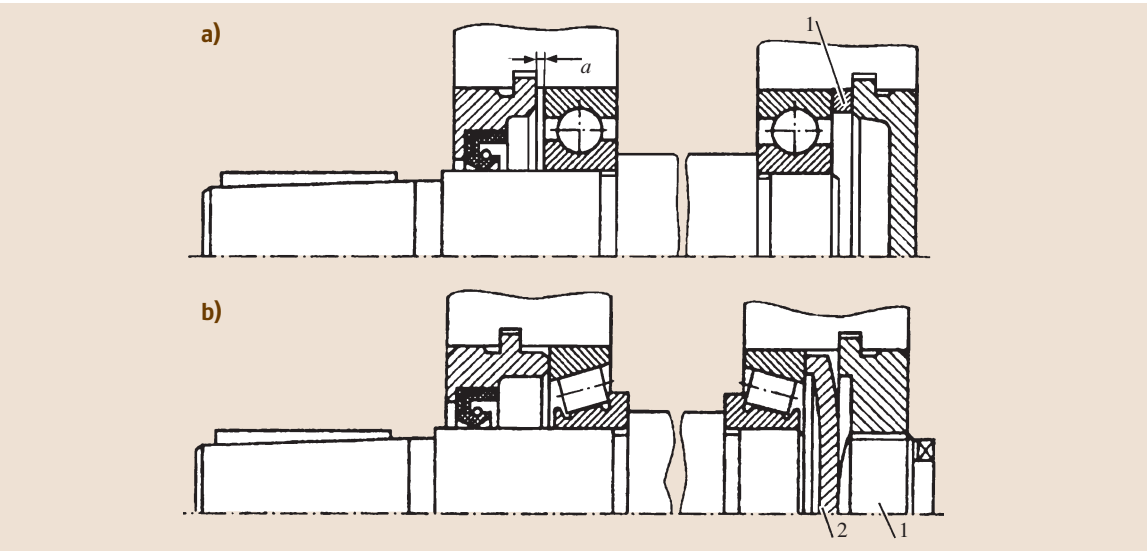


Fig. 6.197a,b Adjustment of clearances in the bearings (diagram 2a). **(a)** With a compensatory ring (1), **(b)** a screw (1), and self-installed washer (2)

fluence (or the value of which is not high). The force applied by the springs must exceed the sum of the axial component from the radial load and the external axial force in radial-thrust bearings.

Supports According to Diagram 2b (Fig. 6.167)

In the axial shaft fixation according to diagram 2b both supports are designed equally according to Fig. 6.199a–f. The axial shaft is fixed with case collars, in which the faces of the outer races are set. The collars for the bearing thrust can be designed in compliance with one of the variants in Fig. 6.185.

The most successful application is represented in Fig. 6.199e. In this variant there are no ledges or grooves in the case. Tapered roller bearings with a ledge on the outer race are currently widely used in the engineering industry.

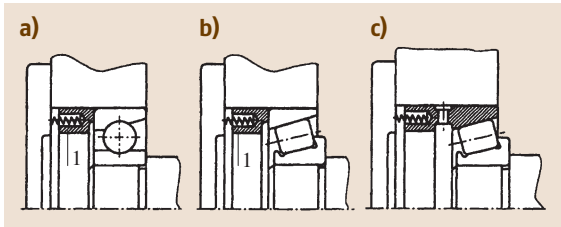


Fig. 6.198a–c Methods to maintain the required stiffness of the supports with the help of a greater number of compression springs installed in **(a,b)** the additional ring, **(c)** in the outer ring of the bearing

Clearance adjustment in the bearings is carried out with the help of the axial displacement of the inner races on the shaft with nuts. It is not necessary to loosen the fit under the moving inner race of the bearing. One nut on one of the shaft ends is enough to adjust the bearings (Fig. 6.200a). If it is additionally necessary to adjust the axial shaft position, nuts are included on both ends (Fig. 6.200b).

The positioning accuracy of the inner races depends on the thread manufacturing accuracy of the shaft and the nut, and the deviation from perpendicularity of the mounting nut face. To increase the accuracy of positioning for the bearings fixed according to diagram 2b (Fig. 6.167) the thread of the shaft is ground in quality products, and the mounting nut face is ground on the threaded arbor.

The firm Timken recommends screwing the adjusting nut (1) onto the shaft thread and locking it at the appointed position (Fig. 6.201). With a ground shaft the mounting nut face (1) is also ground. Some plants provide a permanent axial interference in tapered roller bearings by means of installation of the ring (1) with a large number of springs positioned around the circle (Fig. 6.202a). The firm Game recommends the same solution (Fig. 6.202b).

Mounting of resilient members improves the operating conditions of bearings, as even for this relatively inexact adjustment the axial clearance in the bearings is removed by any thermal shaft extension. Resilient members can be built into the supports not only for tapered roller bearings, but also with ball radial (Fig. 6.202c) and radial-thrust (Fig. 6.202d) bearings.

6.12.4 Design of Shaft Supports of Bevel Pinions

The diagrams of the axial shaft fixation of bevel pinions are shown in Fig. 6.203. In bevel gearing units cantilever fastening of the pinion shaft is widely used (Fig. 6.203a–c). In this case, the structure of the unit is simple, compact, and convenient to assemble and adjust. The disadvantage of the cantilever pinion position is the higher stress accumulation along the tooth of the pinion. If the pinion is positioned between the supports (Fig. 6.203d), stress accumulation is lower as a consequence of a decrease in the shaft flexure and the rotary angle of the section at the installation site of the bevel pinion, but design of the supports according to these diagrams results in a substantial structure meshing of the housing parts and the gear wheel, which is why it is relatively rarely used in practice. The diagram according to Fig. 6.203a (diagram 2b in Fig. 6.167) is primarily applied.

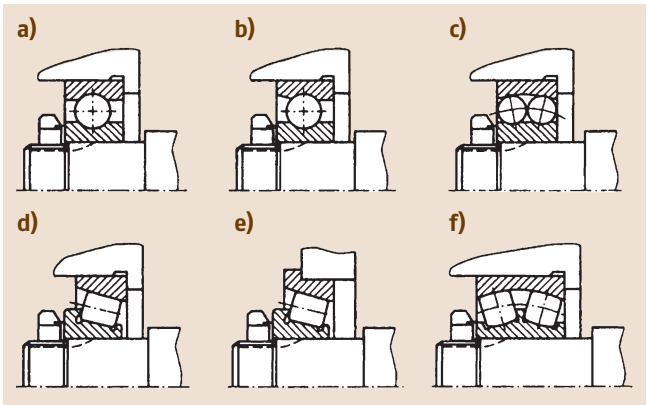


Fig. 6.199a–f Methods of executing the fixing support in diagram 2b using the following bearings: ball radial single-row **(a)**, ball radial-thrust **(b)**, all radial double-row spherical **(c)**, tapered roller **(d,e)** roller radial double-row spherical **(f)**

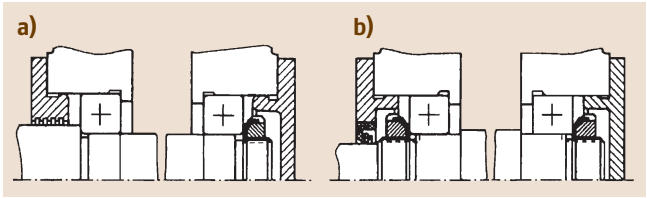


Fig. 6.200a,b Adjustment of clearances in bearings (diagram 2b) **(a)**, with a nut, **(b)** with two nuts

The shafts of bevel pinions are short, so thermal axial deformations do not play such an important role as in long shafts. The distances between the bearings are relatively small, and the forces acting on the shaft and its supports are large. Load accumulation at the cantilever position of the pinion can be decreased by increasing the rigidity of the unit. Higher rigidity requirements are required for high accuracy of the axial position of

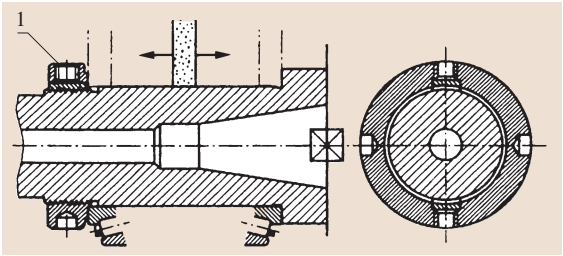


Fig. 6.201 Grinding of the mounting face of the adjustment nut

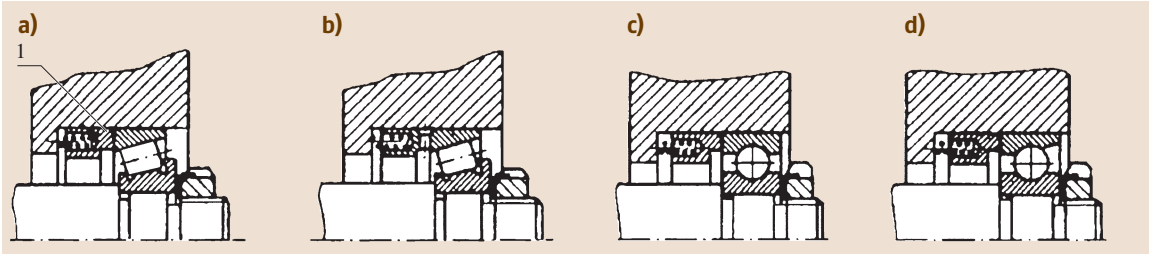


Fig. 6.202a–d Methods of maintenance of the required stiffness of the supports with greater number of compression springs installed in **(a,c,d)** the additional ring (1) **(b)** in the outer ring of the bearing

the bevel pinion, which is necessary according to the working conditions of bevel gearings.

In the unit design, the directions of the tooth dip and of the pinion rotation are chosen equal to the axial force in the toothing, directed from the vertex of the pitch cone of the pinion. In the unit structures of bevel pinions radial-thrust bearings are used, mainly tapered roller ones, because they have greater load-carrying capability and are cheaper, which ensures greater rigidity of the supports. More expensive ball radial-thrust bearings are applied for relatively high rotational frequencies ($n > 1500\,\mathrm{min}^{-1}$) for loss enhancement in the supports, as well as when required for high rotational accuracy.

As indicated, the bearing installation according to the diagram in Fig. 6.203a finds primary application in bevel power trains. Figure 6.204 shows a typical shaft structure of the bevel pinion fixed according to this diagram. The forces acting in the bevel gearing cause a radial reaction of the supports. The radial reaction is considered to be applied to the shaft at the cross point of its axis with the normal lines drawn through the centers of the contact areas on the races of the bearings. Let us designate b as the distance between the reaction points, a as the size of the console, d as the shaft diameter at the installation site of the bearing, and l as the distance from the reaction point of the pinion support closest to the vertex of the pitch point. During design it should be assumed that $d \geq 1.3a$, and that b is the greater of $b \approx 2.5a$ and $b \approx 0.6l$. Engineers aim to obtain the minimum dimension a for reduction of the bending moment acting on the shaft. After this dimension has been determined, the distance b is assumed according to the given ratios. Then the unit is rather compact. The bearing positioned closer to the bevel pinion is loaded with a greater radial force and furthermore takes an axial force. Thus, in a number of structures this bearing will have a larger hole diameter.

The typical shaft structure of the bevel pinion fixed according to the diagram of Fig. 6.203b is given in Fig. 6.205. This installation diagram of the bearing has substantial unit dimensions in the axial direction due to adherence to the required ratio between b and a according to the conditions of stiffness. It is not recommended for use in power trains.

The shaft structure of the bevel pinion fixed in the diagram of Fig. 6.203c is shown in Fig. 6.206. For convenience of adjustment of the axial pinion position the

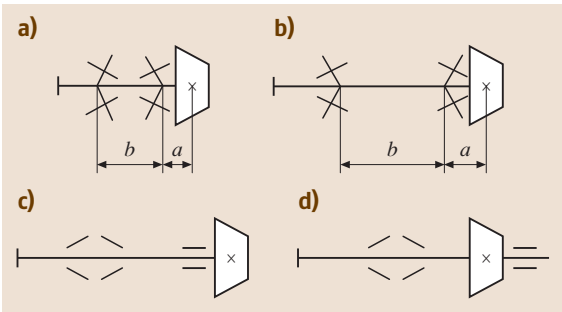


Fig. 6.203a–d Diagrams of the axial fixation of the shafts of bevel pinions. **(a)** Stretched out (diagram 2b), **(b)** end thrust (diagram 2a), **(c,d)** diagram 1b

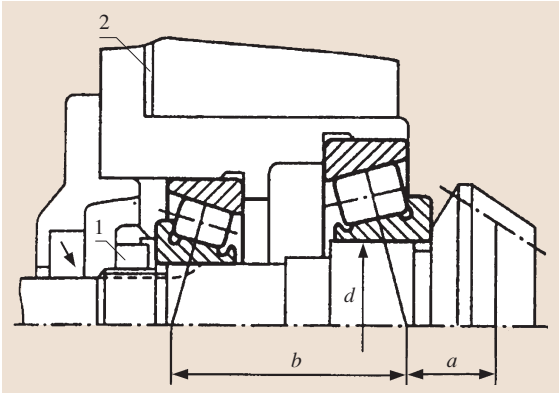


Fig. 6.204 Axial fixation, stretched out (diagram 2b)

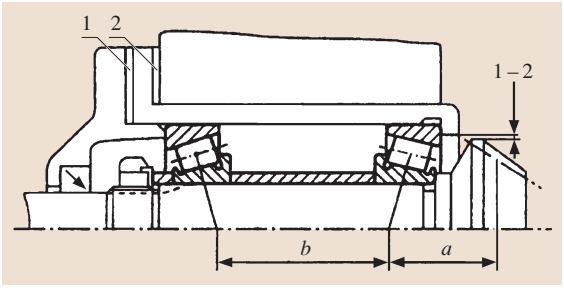


Fig. 6.205 Axial fixation, end thrust (diagram 2a)

fixing support is enclosed in the sleeve. The bearing closest to the pinion is mounted directly in the case hole, which increases the accuracy of the position of the radial pinion.

The floating support can be positioned in one sleeve with the fixing support (Fig. 6.207) by location of the bevel pinion between the supports as shown in Fig. 6.203d in order to simplify the case structure. The disadvantage of structures designed according to this diagram is overdesign of the mating with the pinion bevel wheel.

For the design of the shaft units of bevel pinions, adjustment of the bearing clearances of the fixing supports and of the bevel gearing (of the axial position of the shaft-pinion) is foreseen.

For axial fixation according to Fig. 6.204 the clearances in the bearings are adjusted with a round slotted nut (1), and the axial position of the pinion shaft is adjusted with a package of thin metallic gaskets (2). For axial fixation according to Figs. 6.205–6.207 clearance adjustment in the bearings is carried out with the package of gaskets (1), and the gearing is adjusted with the gasket package (2).

To guarantee the standard (GOST 16984-79) wrench application for screwing, the round slotted nut must be beyond the bounds of the sleeve flanges

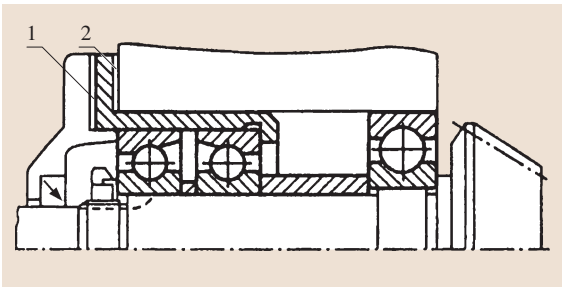


Fig. 6.206 Axial fixation according to diagram 1b with the cantilever position of the pinion

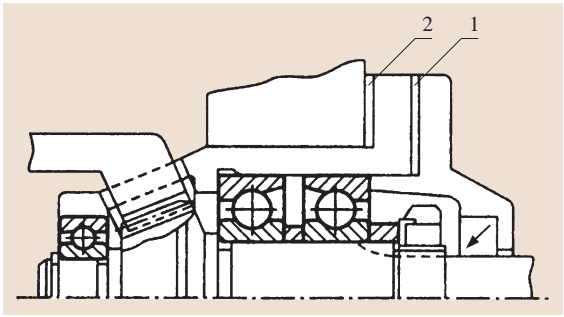


Fig. 6.207 Axial fixation according to diagram 1b with the noncantilever position of the pinion

(Fig. 6.207), which extends the axial unit dimensions and complicates the form of the bearing cap. By application of a nonstandard wrench these disadvantages can be eliminated (Fig. 6.204). To exclude the possibility of damage to the bearing cage by straining with the multitab washer a *joint sealant* can be used to stop the round slotted nut.

Examples of Embodiment of Shaft Units of Bevel Pinions

Figures 6.208 and 6.209 show structures of the bevel pinion input shafts with installation of the bearings according to diagram 2b (Fig. 6.167). Adjustment of the axial clearance in the radial-thrust bearings (Fig. 6.208) is carried out with axial displacement along the shaft with the help of the round slotted nut of the inner bearing race. By adjustment of the toothing the pinion shaft is moved in the axial direction by means of a thickness change in the package of thin metallic gaskets (1) between the case and the flange of the sleeve. In the unit shown in Fig. 6.208a, tapered roller bearings are applied with a thrust ledge on the outer race. The sleeve has a very simple structure.

The bearing positioned closer to the bevel pinion is loaded with a higher radial force and, furthermore, takes an axial force from the toothing side. Thus, in a number of structures this bearing is chosen with a large outer diameter (Fig. 6.208b) or with a large hole diameter (Fig. 6.208c). The bearing is installed directly in the hole of the case. This increases the positional accuracy of the radial pinion.

In the unit of Fig. 6.208d a sleeve with a collar in the hole is used for the positioning of the shaft bearings of the bevel pinion. The installation accuracy of the outer rings in the sleeve depends on the manufacturing accuracy of the collar faces. The presence of the collar in the hole of the sleeve complicates its machining.

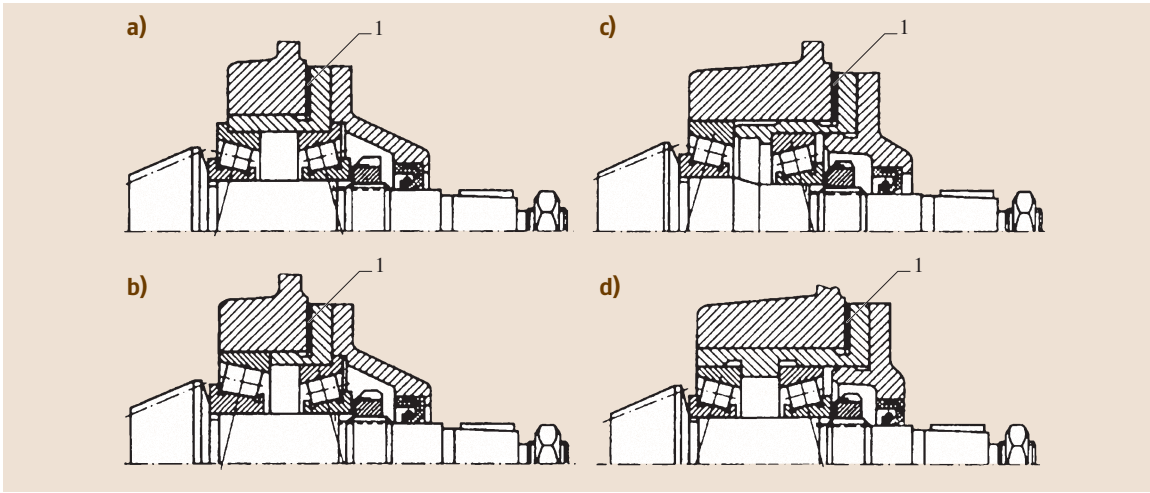


Fig. 6.208a–d Shaft structures of bevel pinions using tapered bearings. **(a)** With a shoulder on the outer race, **(b)** of different dimension series, **(c)** of different inner diameters, and **(d)** mounted in the sleeve with a collar

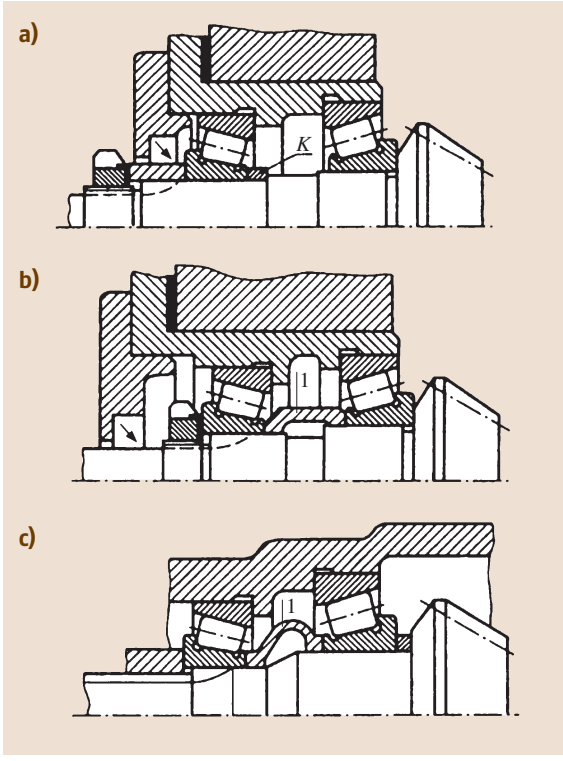


Fig. 6.209a–c Forming the axial clearance in the supports of shafts of bevel pinions. **(a)** With a compensatory ring, **(b)** with a stiff, and **(c)** compliant bushing

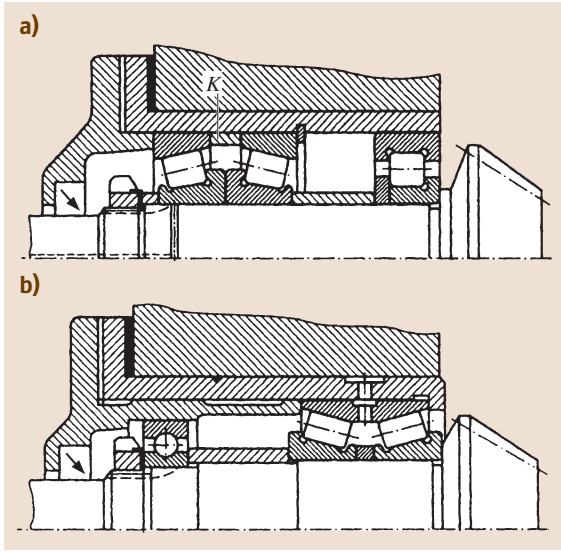


Fig. 6.210a,b Axial fixation of shafts of bevel pinions according to diagram 1b with position of the fixing support **(a)** at the outlet of the shaft, and **(b)** next to the pinion

A peculiarity of the described structures of the sleeves is the fact that their position in the case is not determined by an external cylindrical surface, but rather by a developed flange. This is why the cylindrical surface used only for centering can be substantially reduced (Fig. 6.208c).

Figure 6.209a,b shows unit structures of bevel pinions used in cars (according to information from the firm SKF). The inner race of the left bearing is tightened with a nut up to the stop to the face of the compensatory ring K or to the face of the compensatory bushing (1), which improves its positioning. In some structures there is a compliant steel bushing (1) between the faces of the inner bearing races (Fig. 6.209c). The required preload of the bearing is applied with a torque spanner by the application of a tightening torque determined based on experience.

Figures 6.210 and 6.211 show the structures of the input shafts of bevel pinions with a fixing support and a floating one (diagram 1b, Fig. 6.167). For convenience of adjustment of the axial pinion position both shaft supports, fixing and floating, are enclosed in the sleeve (Fig. 6.210a). Adjustment of the bearing of the fixing support is carried out by means of matching and grinding of the compensatory ring K. Sometimes (Fig. 6.210b) the fixing support is not located at the output shaft as usual, but near the bevel pinion.

In has been noted that a noncantilever position of the pinion is more efficient. However, such structures are more complicated. An additional support can be positioned in a special inner wall (Fig. 6.211a,b). Because the teeth of the bevel pinion are cut on the shaft, the bore diameter for the bearing is small. The mating wheel of the bevel gearing limits the radial dimensions of this support. The fixing support in Fig. 6.211a is adjusted with a cap (1) screwed into the sleeve. In Fig. 6.211b it is adjusted by matching and grinding of the compensatory ring K. The bevel meshing is adjusted with the package of metallic gaskets (2) installed under the flange of the sleeve.

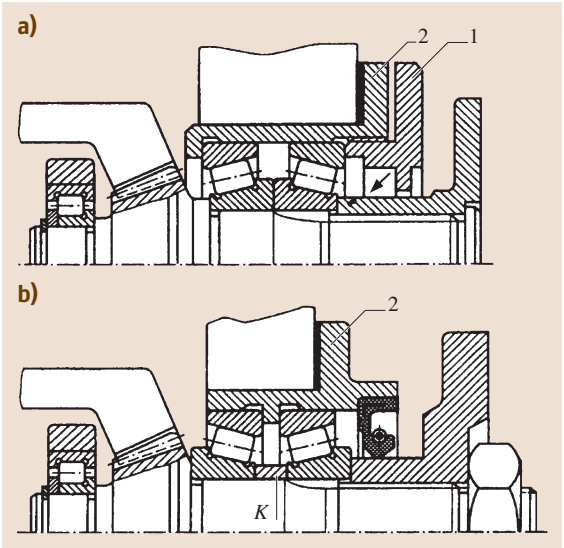


Fig. 6.211a,b Axial fixation of shafts of bevel pinions according to diagram 1b with adjustment of the tapered roller bearings of the fixing support with **(a)** a cap (1), **(b)** and a compensatory ring K

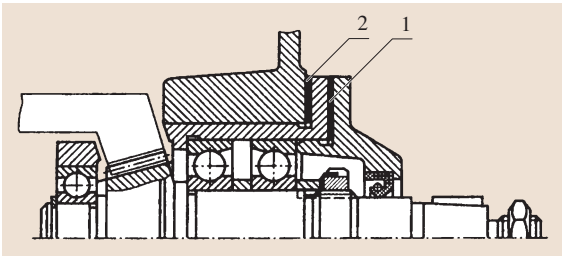


Fig. 6.212 Axial fixation of shafts of bevel pinions according to diagram 1b with adjustment of the ball radial-thrust bearings of the fixing support with metallic gaskets

There is a possible variant of the structure with extra support in the sleeve, as shown in Fig. 6.212. In this case, rigidity of the unit is rather high and for rotation loss enhancement ball radial-thrust bearings can be used in the fixing support, and the radial bearing can be used in the floating support. Bearing adjustment of the fixing support is carried out by means of thin metallic gaskets (1), and adjustment of the bevel meshing is applied with metallic gaskets (2).

6.12.5 Support Design of Worm Shafts

Diagrams for axial fixation of the worm shafts are given in Fig. 6.213. Fixation from the axial displacements according to diagram 2a (Fig. 6.167) is used with the expected differential temperature of the worm and the case being up to 20 °C and relatively short shafts (Figs. 6.213a and 6.214). Thus for shaft installation $d = 30–50$ mm and in ball radial-thrust bearings the ratio l/d is not more than 8, whereas in tapered roller bearings l/d is not more than 6.

Because the substantial axial force influences the worm, radial-thrust bearings are mounted in the supports. Mainly tapered roller bearings are used

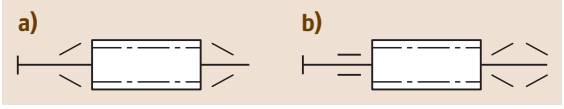


Fig. 6.213a,b Diagrams of the axial fixation of worm shafts. **(a)** Thrust (diagram 2a), **(b)** diagram 1b

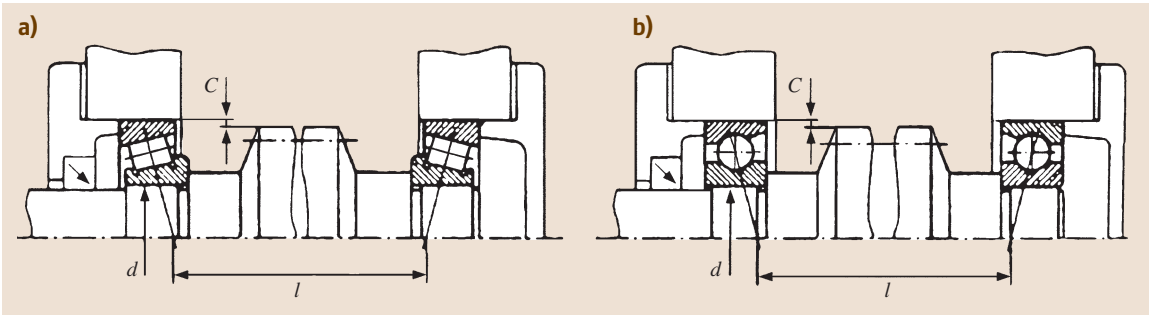


Fig. 6.214a,b Axial thrust fixation using (a) tapered roller bearings, (b) ball radial-thrust bearings

(Fig. 6.214a). Ball radial-thrust bearings are used in continuously operating gearings in order to reduce power loss and maintain heat generation in the supports, as well as for adherence to the manufacturing accuracy of the unit components (Fig. 6.214b). However, the dimensions of supports designed using radial-thrust bearings are larger than those of tapered roller bearings as a consequence of their lower load rating. Thus, the final choice of the shaft supports of the worm is sometimes made after comparative calculations and tracings. It should be taken into account that it is not recommended to install radial-thrust bearings with a large contact angle ($\alpha > 18°$) according to diagram 2a. The diagram with a fixing support and a floating support (Fig. 6.213b) is used because of the application requirements of these bearings, as well as the large expected thermal deformations of the shaft for fastening in the case of the worm shaft.

Figure 6.215 shows the most common variant of the fixing support of worm shafts. Due to the high axial force acting on the worm shaft, radial-thrust bearings, i. e., tapered roller bearings or ball bearings with a large contact angle, are used in the fixing support. Because radial-thrust single-row bearings are subjected to an axial force from only one direction, in the fixing support two such bearings are mounted for shaft fixation in both directions.

The clearances in the bearings of the fixing support are adjusted with the package of thin metallic gaskets (1), which are installed under the flange of the bearing cap. Instead of the adjusting gaskets a precisely longways fitting ring K (Fig. 6.215) is sometimes mounted between the outer bearing races.

A radial-thrust double bearing in combination with the radial one is used for substantial axial loads in the fixing support. Some structures of such supports are given in Fig. 6.216a,b. Installation of thrust bearings on the horizontal shaft is undesirable as the axial force loads one of the last rings and unloads another. In con-

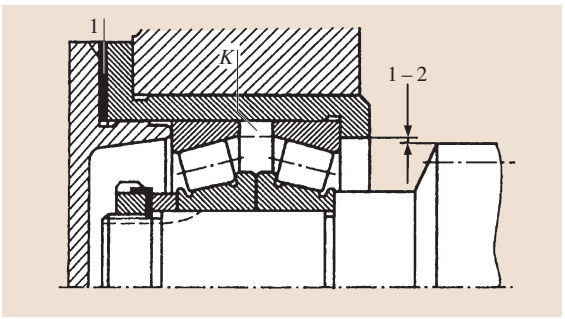


Fig. 6.215 Fixing support of a worm shaft using two tapered roller bearings

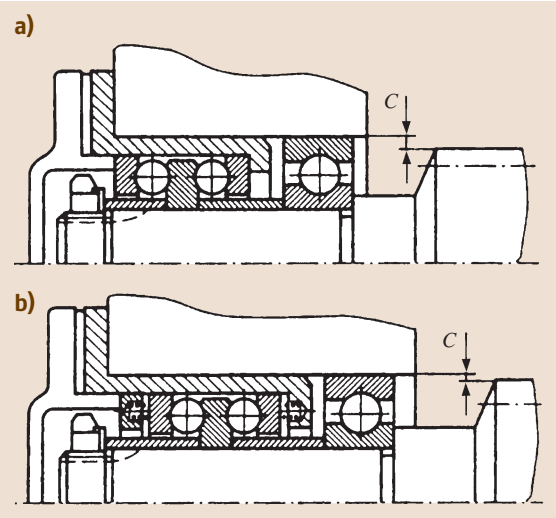


Fig. 6.216a,b Fixing support of a worm shaft using a ball radial single-row bearing and a ball thrust double bearing, the races of which are (a) not tightened, (b) and tightened with compression springs

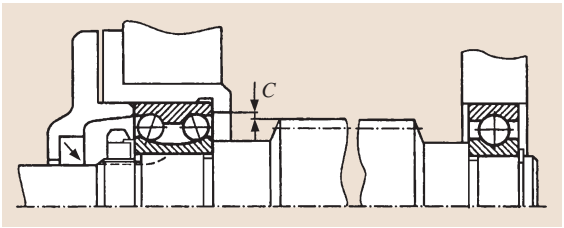


Fig. 6.217 Axial fixation of a worm shaft according to diagram 1a using a ball double-row radial-thrust bearing in the fixing support

tact with the unloaded ring the balls creep under the action of inertial forces (the gyroscopic effect). This results in greater heat in the bearing and faster fracture. To avoid this increased creep the races of the thrust bearings are tightened with springs (Fig. 6.216b).

The axial fixation according to diagram 1a (Fig. 6.167) is seldom used. Figure 6.217 shows the structure of worm-shaft supports engineered by SKF. In the fixing support a very complicated and expensive ball radial-thrust double-row bearing is applied. To insert the worm shaft with the bearings into the sleeve or the case a clearance of $C \geq 1-2$ mm is foreseen (Figs. 6.214–6.217).

Examples of the Embodiment of Worm-Shaft Units

The minimum support dimensions in the radial direction, as well as the minimum distance between the bearings, can be obtained by installation of combined radial-thrust needle bearings (Fig. 6.218, according to information from the firm NADELLA, France). To position the flank of the combined needle bearing, housing parts must be machined. The seal on the input shaft end is positioned in a smooth hole, which is intended for installation of the bearing. The required clearance for the bearing operation is ensured with

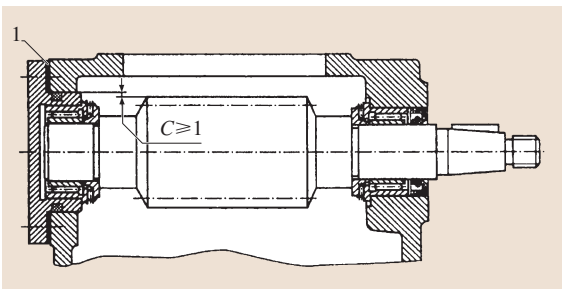


Fig. 6.218 Axial fixation of a worm shaft using combined radial-thrust needle bearings

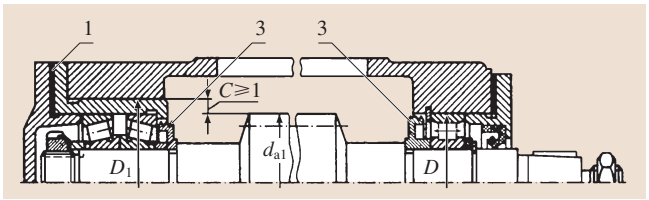


Fig. 6.219 Axial fixation of a worm shaft according to diagram 1b using roller bearings

the help of the metallic gaskets (1). Russian industry also produces similar bearings. On assembly the worm shafts are inserted into the case through the hole intended for installation of the bearings. Thus, the diametral dimensions of the worm or the components positioned on the shaft must be less than the hole diameter $2C$. If the worm diameter d_{a1} is more than the bearing diameter D, the bearing is mounted in the sleeve (Fig. 6.219).

Figure 6.219 shows an embodiment of the worm-shaft unit by mounting of the bearings according to diagram 1b (Fig. 6.167): the left support is fixing, the right one is floating. With such an installation configuration of the bearings the fixing support can support substantial axial forces, because tapered bearings with a large cone angle can be applied.

Possible variants of the fixing support of the worm shaft are given in Fig. 6.220. In Fig. 6.220a for fastening of the bearings a thrust collar is foreseen in the case, which however complicates machining of the mount-

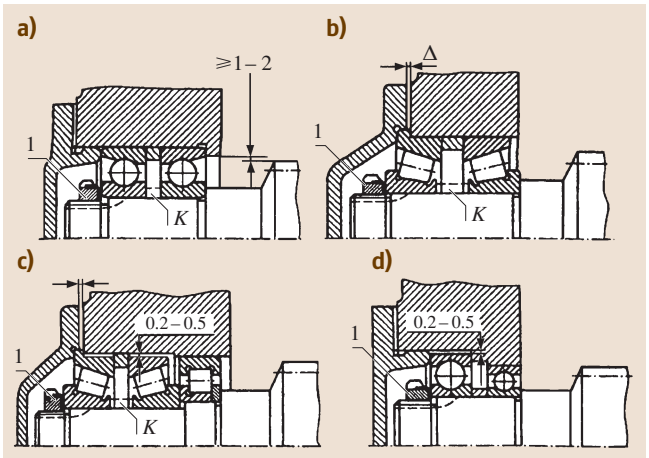


Fig. 6.220a–d Variants of design of the fixing support using **(a)** a thrust collar in the case, **(b)** a bearing with a thrust shoulder, **(c)** a roller, and **(d)** and ball bearings taking radial and axial loads separately

ing holes for the bearings. Application of bearings with a thrust ledge on the outer race (Fig. 6.220b) substantially simplifies a structure: there is a smooth hole in the case and there is no sleeve. In Fig. 6.220b tapered roller bearings are oriented towards each other with the wide faces of the outer races, and in Fig. 6.218 with the wide faces outwards. The variant shown in Fig. 6.220b is characterized by high angular rigidity.

The diametral dimensions of the support can be reduced as necessary if different bearings support the radial and axial forces. In the structure in Fig. 6.220c tapered roller bearings are mounted in the case with little clearance and therefore can take only axial force. By unloading the tapered bearings from the radial force their lifetime can be increased. The radial force is supported by the radial bearing with short cylindrical rollers. In order to bear the radial load radial bearings of other classes can be used. On the whole, the support in Fig. 6.220c is more expensive, e.g., than the support in Fig. 6.220b.

Figure 6.220d shows a structure of the fixing support of the worm, where ball, radial, and radial-thrust bearings with a split inner race, are applied. Here, as in Fig. 6.220c, clearance is included between the mounting hole and the bearing, so that the radial-thrust bearing only supports an axial force. The radial-thrust bearing is of the nonadjustable class; the required axial clearance is provided by production of the bearing. In other variants (Fig. 6.220a–c) the bearings of the fixing support are adjusted with a nut (1). Then the precisely fitting rings K (shown in the figures by hatch lines) are sometimes mounted between the races of the bearings. Attention should be paid to the fact how the bearing caps are installed in Fig. 6.220b,c. By tightening the fastening bolts the cap tightens a ledge on the outer bearing race to the case. A small clearance Δ must necessarily be present between the cap face and the mount of the case. This fastening guarantees transmission of axial forces in any direction from the bearing to the case.

6.12.6 Supports for Floating Shafts

Shafts are floating if both their supports are floating. In this case, the possibility of self-installation of the floating shaft relative to another shaft fixed from axial displacements is secured. This self-installation is required, e.g., in herring-bone or helical gearings representing a separated chevron. In the manufacture of the wheels of these gearings an error in the angular tooth position of one semichevron relative to the tooth of another semichevron is unavoidable. Due to this error first the teeth of only one semichevron engage. The axial force arising in the toothing works to move the wheel with the shaft along the shaft axis. If the supports allow, the shaft moves to such a position where the teeth of both semichevrons engage, and the axial forces arising in them are balanced. In this case, the axial fixation of the shaft is not carried out in the supports but by the teeth of the chevron gears.

Radial bearings are used as supports for floating shafts. Bearings with short cylindrical rollers are mostly applied. The construction diagrams shown in Fig. 6.221 are the most common.

Diagram in Fig. 6.221a

The inner races of the bearings are fastened on the shaft, and the outer ones are in the case. Axial floating of the shaft is ensured by the fact that the inner bearing races with a roller set can shift in the axial direction relative to the fixed outer races. Axial floating of the shaft occurs during its rotation. Then the force required for its displacement is very small, which is an advantage of this configuration.

Disadvantages of this approach are as follows:

- The necessity to apply very rigid shafts and guarantee of a high coaxiality grade of the mounting shaft and case surfaces as a consequence of the high sensibility of these bearings to race warps.

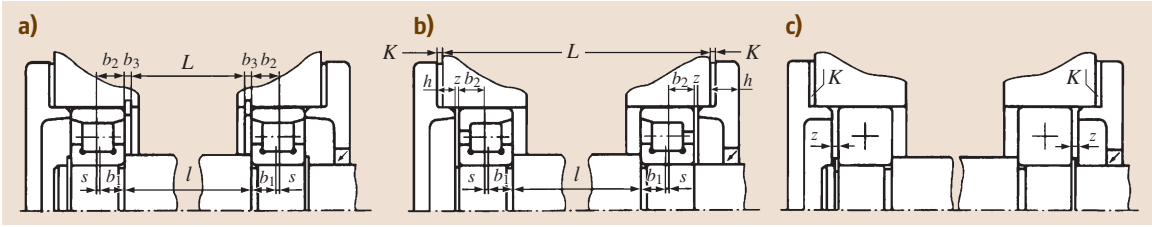


Fig. 6.221a–c Analytical models of floating shafts with installation of bearings with (a,b) short cylindrical rollers, (c) of other classes of radial bearings

- The possibility of substantial initial (after assembly) axial race displacement s, which is not compensated later. Errors in the dimensions l, L, b_1, and b_2 cause this displacement and also the fact that the axial shaft position depends on the axial position of the engaged wheels, which has an accidentally wide spacing in values.
- The need for comparatively precise production of the components according to the dimensions L and l. These dimensions (shown in Fig. 6.221a) along with other dimensions form a dimensional chain. The errors in the component manufacture according to these dimensions result in axial displacement of the bearing races.

Diagram in Fig. 6.221b

The outer races have some freedom from axial displacement. Displacement into the case is restricted to the ledges of both bearing races; towards the bearing caps it is restricted by a clearance z. The value of the clearance $z = 0.5–0.8$ mm depends on the unit dimensions and manufacturing accuracy of the teeth of the mated chevron gears, and their assembly accuracy.

With axial floating of the shaft the inner races of the bearings with roller sets shift relative to the outer races. At the start of axial shaft floating the rollers of the bearings displace the outer races towards the caps in such a way that the races find their place and are fixed later.

The advantages of this configuration are the following:

- Easy shaft floating because of the low axial force.
- The possibility of adjustment of the initial value of s – the axial displacement of the races – to the minimum. This is achieved by means of matching of the compensatory gaskets K mounted under the flanges of both bearing caps.
- The production of the components according to the dimensions l, L, and h in compliance with free tolerances (e.g., of accuracy degree 14). Possibly accumulated errors are eliminated with the compensatory gaskets K.
- The absence of stops for the outer bearing races in the case holes, which makes their machining easier.

The disadvantage of this configuration, as with the previous one, is that its application is limited to stiff shafts and high manufacturing accuracy of both the shafts and the case holes.

Diagram in Fig. 6.221c

In this configuration, in the supports radial ball single-row, ball, or roller double-row spherical bearings are applied. The choice of one or another bearing class is defined by the required load rating and shaft stiffness.

The inner bearing races are fastened onto the shaft, whereas the outer races are free and can move along the holes of the case. The displacement value is restricted by the clearances z set on assembly by matching the compensatory gaskets K. Axial shaft floating, if its value is not more than the axial clearance in the bearings, occurs at the expense of this clearance relative to the fixed outer bearing races. If the axial shaft displacement exceeds the axial clearance in the bearings, by floating of the shaft the outer bearing races slide in the holes of the case, which results in wear of the hole surface. To decrease this wear tempered-steel bushings are sometimes placed into the holes of the case.

The advantage of this configuration is than it can be applied for nonrigid shafts and low coaxiality grade of the mounting surfaces of the shaft and the case. The absence of stops for the outer bearing races in the holes of the case can also be considered an advantage.

The disadvantages of this configuration are the following:

- The presence of kinetic friction of the outer bearing races along the holes of the case.
- The necessity of the application of substantial axial force for realization of the shaft floating.
- The use of tempered-steel bushings makes the supports more expensive and reduces the positioning accuracy of the shaft.

Examples of the Embodiment of Floating Shaft Units

Figure 6.222 shows structures of the input shafts of a single-reduction gear unit with chevron gears made according to the configuration shown in Fig. 6.221a,b. The shafts are floating. The axial position of the floating shaft is determined by the teeth of the semichevrons, which are inclined in different directions. The conjugated shafts are fixed relative to the case.

The outer race of the bearing without ledges (Fig. 6.222a) is tightened with a face of the clamp-on cap to ring (1). This ring can be solid if the jointing plane of the case goes through the shaft axis. If the case is made without a split, (1) is a spring planar thrust inner ring. In the floating support shown in Fig. 6.222a it is recommended to fasten the inner bearing race from

two sides to prevent it from accidentally coming off the shaft. For compensation of unavoidable manufacturing inaccuracy the compensatory ring (3), the thickness of which is matched at assembly, is mounted for the length of the components between the spring ring (2) and the face of the inner bearing race.

By application of a bearing with one ledge on the outer race (Fig. 6.222b) the required axial position of the clamp-on caps is set on assembly by matching the thin metallic gaskets (4). The outer races are free from axial displacement for the value of the clearance z in the direction of the bearing cap. There is no need to fasten the inner bearing race onto the shaft.

At the start of axial shaft floating the rollers of the bearings displace the outer races by a certain value in the direction of the caps. Then the clearance z reduces and later, at the expense of thermal deformations of the shaft, it is completely removed. After the races find their position they become fixed (Fig. 6.222c). Then there is an axial clearance s between the rollers and the ledge of the outer race by the shaft floating. The clearance s changes during the working process over certain ranges, which are determined by the manufacturing accuracy of the teeth of the gear wheels. An important advantage of this configuration is the possibility of adjustment of the initial value of the axial displacement of the outer and inner bearing races.

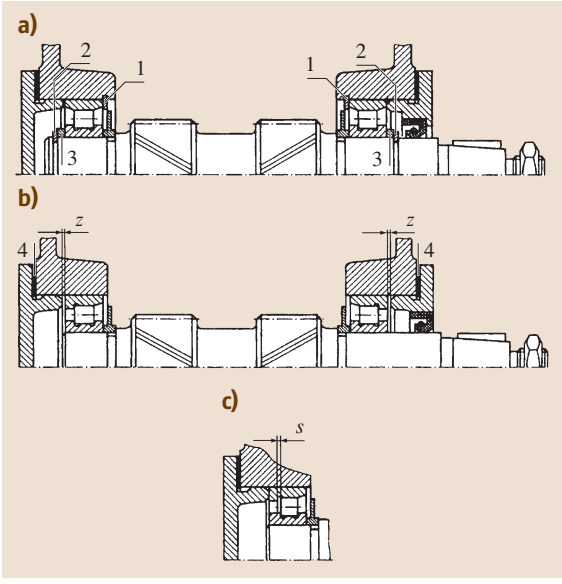


Fig. 6.222a–c Structures of the floating shafts of herringbone gears using bearings with short cylindrical rollers: **(a)** without shoulders on the outer race, and **(b)** with one-shoulder outer race. **(c)** Diagram of the axial clearance in the bearing

6.12.7 Supports for Coaxial Shafts

These supports are designed, e.g., in the coaxial cylindrical double-reduction gear unit (Fig. 6.223), as well as in multiengine transmissions. In this case, on the inner case wall, bearings for the coaxial shafts (1 and 2), which have different overall dimensions, are located side by side. One of them is a support for the high-speed shaft, while another is a support for the low-speed one. As a rule, the shafts are fixed according to diagram 2a (Fig. 6.167). Figure 6.224 shows design versions of the support of coaxial shafts (remote element A, Fig. 6.223).

In Fig. 6.224a the holes for the bearings are made directly in the inner wall of the case. Machining of the holes is carried out from both sides, forming the collars for the bearings in both holes. This presents some machining problems; however with this design the greatest installation accuracy of the bearings can be achieved.

Boring of the hole can be simplified if it is done with a through diameter of D_2 (on the outer diameter of the bigger bearing, Fig. 6.224b). However, for installation of a bearing with a smaller outer diameter D_1 an additional component is used: the ring (3), which is fixed with a collar on the outer surface which goes into the groove of the split frame. The bearings are inserted up to the stop into the faces of the ring (3), which is why the manufacturing accuracy of the ring must be high. Thus simplified hole boring is achieved with the application of ring (3), execution of the groove in the case,

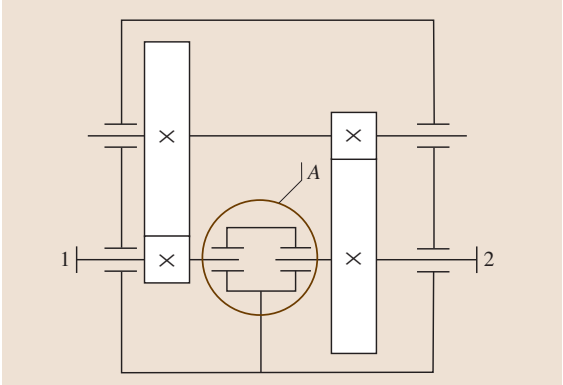


Fig. 6.223 Diagram of the coaxial cylindrical double-reduction gear unit

and the required use of the removable cap in the inner wall of the case.

In the version shown in Fig. 6.224c the ring (3) does not have a fixing flange and, therefore, the groove is not needed in this case. The structure of the ring is simpler and machining of the case hole is also simpler. However, here the shafts (1 and 2) form a general system. Adjustment of the axial clearance for four bearings of both shafts is carried out simultaneously. The main disadvantage of this version is that the axial forces acting on one shaft load the bearings of another shaft.

Depending on the version the points below can be followed:

- For the version shown in Fig. 6.224a the positional accuracy is higher, as there is no additional component with corresponding errors and no additional conjugation with the case.
- The version in Fig. 6.224c should be preferred to that in Fig. 6.224b as it is simpler and more economical due to the installation of the ring (3) when using radial ball bearings in the supports. It goes without saying that, by matching the bearings, the axial forces acting on both shafts (1 and 2) should be taken into account.

Examples of the embodiment of the supports of coaxial shafts are presented in Fig. 6.225 with a stop of the bearings into the collars either of the case (Fig. 6.225a) or the ring with flanges (Fig. 6.225b).

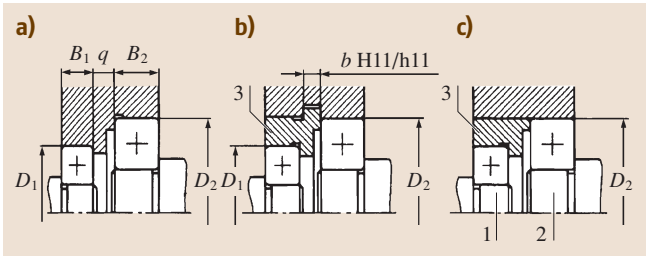


Fig. 6.224a–c Variants of the support design of coaxial shafts. **(a)** Holes for the bearings are made directly in the case. **(b)** A ring with a ledge is used. **(c)** A ring without a fixing ledge is used

6.12.8 Lubrication of Bearings

Lubrication of bearings is carried out with the help of semisolid lubricants and liquid oils. In some cases solid lubricants are used. The choice of the lubricant type depends on operating conditions and, mainly, on the bearing temperature, rotational frequency, acting loads, and the structure of the bearing and the bearing unit.

For lubrication of rolling bearings running under normal conditions semisolid lubricants are predominantly used. These have the following advantages in comparison with oils: they do not need compound sealing devices, they have higher corrosion protection characteristics, they are more economical, they are retained better in the bearing unit, especially for an inclined or a vertical position of the shaft, and they better protect the bearing from moisture penetration and contamination from the environment. The service life of semisolid lubricants often exceeds the lifetime of the bearing, so the bearings do not require repeated lubrication.

However, application of liquid lubricants allows one to reduce the frictional moment and increase the limit rotation frequency by factor of 1.2–1.5. With the help of liquid lubricants, heat elimination and removal of wear debris occur.

Solid lubricants are applied for bearings running in conditions for which liquid and semisolid lubricants are unusable (vacuum, high and low temperatures, hostile environments, radioactive emission, equipment in the food and textile industries, and optical systems).

Semisolid lubricants consist mainly of a liquid base, thickener, and additives to improve their operational properties. The thickener, which comprises 8–25% of the total mass of the lubricant, forms a three-dimensional grid in the form of a fibrous frame whose cells retain oil. This is why, with low loads, the

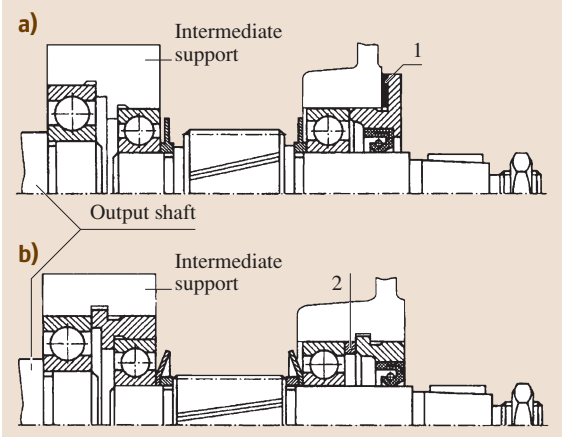


Fig. 6.225a,b Structures of the support of coaxial shafts: with stopping of the bearings **(a)** into the collars of the case, **(b)** into the collars of the ring with a ledge

semisolid lubricant behaves like a solid; it does not spread under its own weight and is retained on tilted and vertical surfaces.

Lubricants with calcium and lithium thickeners are used for bearings. Mineral and synthetic oils with a kinematic viscosity of 15–500 mm^2/s at 40 °C are applied as a dispersion medium. For the lubrication of rolling bearings semisolid lubricants of classes 2 (predominantly) and 3 according to the National Association of Lubricating Grease Institute (NLGI) standards are recommended. Most often lubricants with a lithium base are applied, which are resistant to water and are corrosion protected.

One type of lubrication method has a permanent quantity of lubricant that is intended for the entire lifetime of the bearing. The other requires periodic addition and changes of the lubricant. In the first case, the lifetime of the lubricant is equal to or greater than the lifetime of the bearings or the maintenance cycle of machines with built-in bearings. Closed bearings filled with a lubricant on manufacture with safety washers or with contact seals belong to this class. Bearings with built-in safety washers are applied in units where contamination is not high and water, vapor, etc., do not penetrate, or in units where the absence of friction in this noncontact seal in the case of high rotational frequencies or high temperatures is important.

Bearings with built-in contact seals are applied in units where it is impossible to ensure an external seal due to a lack of space, where the possibility of contamination is normal and ingress of moisture is possible, or if it is necessary to guarantee a long lifetime without maintenance.

As a liquid lubricant refined mineral (petroleum) oils are mostly used for bearings. Liquid synthetic oils (diether, polyalkylen-glycol, fluorine-carbonic, silicone) in comparison with mineral oils demonstrate better stability, viscosity, and pour point. They are used at high or low temperatures, and high rotational frequencies.

The choice of lubricating oil is determined by the viscosity required to ensure effective lubrication at the operating temperature. The dependence of the oil viscosity on the temperature is characterized by the viscosity index (VI). A higher VI indicates less viscosity dependence on temperature. The wider the range of operating temperatures, the greater the viscosity index of the oil used should be. For lubrication of rolling bearings oils with VI of 85 and higher should be used.

Table 6.94 shows a classification of kinematic viscosities in accordance with the recommendations ISO 3448.

Table 6.94 Classification of kinematic viscosities in compliance with ISO 3448

Viscosity class	Kinematic viscosity (mm^2/s) at 40 °C		
	Average	Minimum	Maximum
ISO VG 2	2.2	1.98	2.42
ISO VG 3	3.2	2.88	3.52
ISO VG 5	4.6	4.14	5.06
ISO VG 7	6.8	6.12	7.48
ISO VG 10	10	9.00	11.0
ISO VG 15	15	13.5	16.5
ISO VG 22	22	19.8	24.2
ISO VG 32	32	28.8	35.2
ISO VG 46	46	41.4	50.6
ISO VG 68	68	61.2	74.8
ISO VG 100	100	90.0	110
ISO VG 150	150	135	165
ISO VG 220	220	198	242
ISO VG 320	320	288	352
ISO VG 460	460	414	506
ISO VG 680	680	612	748
ISO VG 1000	1000	900	1100
ISO VG 1500	1500	1350	1650

To increase the performance characteristics of the oil various additives are used. The most common additives are antioxidants, anticorrosives, antifoams, antideterioration, and antiscoring.

Preference is given to oil used in the conjugate units (bearings and gear wheels are usually lubricated from a common oil reservoir). The use of oil with higher viscosity is advisable in the case of high loads and low velocities.

Efficiency of lubrication depends on the degree of separation of contact surfaces by the lubrication layer. To form an appropriate layer the lubricant must have a certain minimum viscosity, ν_1, at the operating temperature. The value of the minimum required kinematic viscosity ν_1 can be determined from the nomogram shown in Fig. 6.226, depending on the mean diameter d_m (mm) of the bearing and its rotational frequency n (min^{-1}). This nomogram corresponds to the results of the latest research in the field of the tribology of rolling bearings.

If the operating temperature of the bearing is known from field experience, or can be determined by other means, the kinematic oil viscosity ν at the base temper-

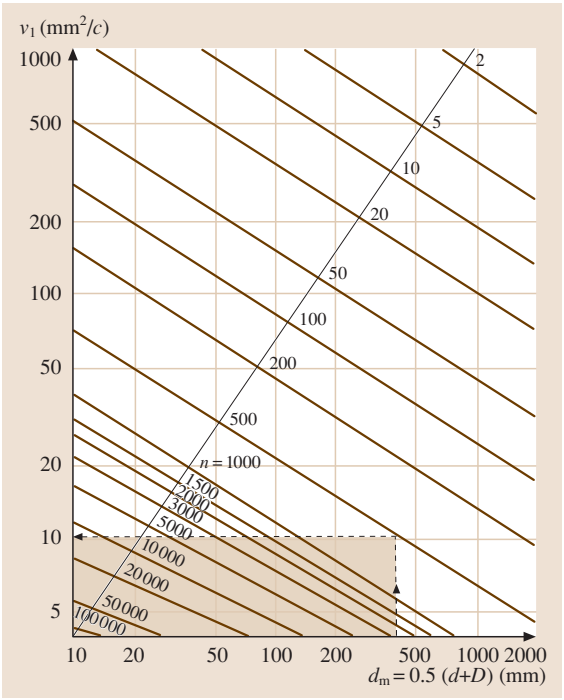


Fig. 6.226 Chart for the determination of the initial viscosity ν of oil that ensures the required viscosity ν_1 at the operating temperature t

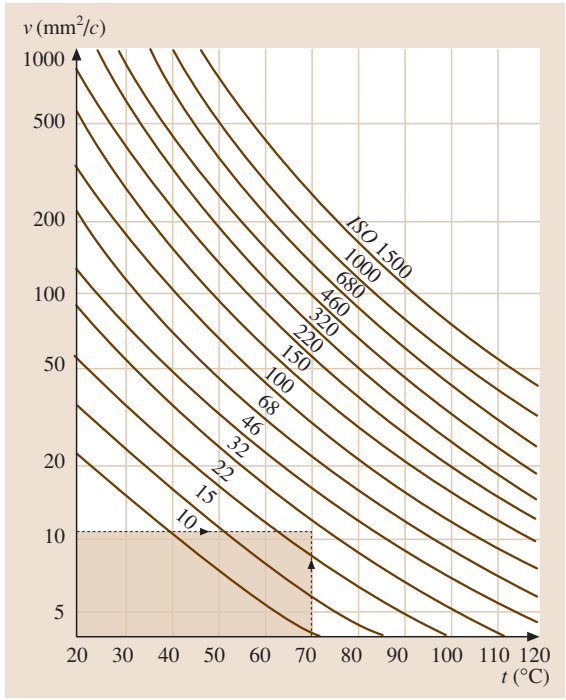


Fig. 6.227 Nomogram for defining kinematic oil viscosity ν of the base temperature 40 °C

ature of 40 °C can be determined from the nomogram shown in Fig. 6.227, depending on the required viscosity ν_1 and the operating temperature of the bearing. This nomogram is correct for oils with a viscosity index of VI 95. The higher the viscosity ν is at the base temperature, the higher the lifetime of the bearings, but with higher viscosity the operating temperature of the bearings increases.

The lubrication rate can be indirectly estimated by using the parameter of relative viscosity: $K = \nu/\nu_1$. With the relative viscosity $K < 1$ oil with antiscuffing additives (EP) is recommended, and with $K < 0.4$ the application of such oil is obligatory.

From the nomogram in Fig. 6.227 the kinematic oil viscosity ν at a base temperature of 40 °C can also be determined depending on the required viscosity ν_1 at the operating temperature of the bearing, i. e., the necessary viscosity class of oil can be determined.

Example

The viscosity class of oil for lubrication of a bearing with a hole diameter of $d = 340$ mm and an outer diameter of $D = 420$ mm, which is to run at the rotational frequency $n = 500\,\text{min}^{-1}$ and at an operating temperature of $t = 70$ °C, is to be determined.

Solution

The mean diameter of the bearing is $d_m = 0.5(d + D) = 0.5(340 + 420) = 380$ mm. From the nomogram in Fig. 6.226 it is determined that, for $d_m = 380$ mm and $n = 500\,\text{min}^{-1}$, the minimum kinematic oil viscosity ν_1 required for effective lubrication at the working temperature of the unit must not be lower than 11 mm²/s. From the nomogram shown in Fig. 6.227 we determine that, for a bearing operating at a temperature of 70 °C, oil with a viscosity of not less than 29 mm²/s at the standard temperature of 40 °C is necessary. Oil of the viscosity class ISO VG 32 is chosen (Table 6.94), the kinematic viscosity is $\nu = 32$ mm²/s at a temperature of 40 °C. Then from Fig. 6.227 the viscosity of the set oil at an operating temperature of 70 °C is determined as $\nu = 12.5$ mm²/s, and the relative viscosity parameter is calculated as $K = \nu/\nu_1 = 12.5/11 = 1.14$.

For most bearings with average dimensions (except roller spherical, tapered, and roller thrust bearings) that run under normal conditions oils with a kinematic viscosity of $\nu = 12$ mm²/s at working temperature are

recommended. For roller tapered and spherical bearings one uses $\nu = 20\,\mathrm{mm^2/s}$, and for roller thrust bearings one uses $\nu = 30\,\mathrm{mm^2/s}$. Oils with a viscosity of less than $12\,\mathrm{mm^2/s}$ are applied for high-speed compact bearings, especially when low starting moments are needed.

The most common oil supply methods in bearing units are the following: oil reservoirs, wicks and spraying, spiral grooves, cone nozzles, metering oilers, continuous flushing, periodic injection, oil mist, and air–oil mist.

For bearings running at moderate rotational frequencies and horizontal installation of the shaft the simplest lubrication methods are used, i. e., spray and oil reservoirs. In the latter case oil is filled into the case in such a way that its level is located in the center of the lower ball or roller (for a rotational frequency of up to $3000\,\mathrm{min^{-1}}$); for higher frequencies it is placed slightly lower.

Solid lubricants are used in the form of powders and thin coatings, or in the form of self-lubricated structural materials for the manufacture of retainers. Molybdenum disulfide, tungsten disulfide, graphite, fluorocarbon polymer, as well as composites made on their basis are the most often used solid lubricants. Solid lubricants are produced in the form of powders and pastes, or are colloid-dispersed or suspended in liquids and added to the lubricant materials or directly coated onto the components of the bearings, as well as in the form of briquette used for the manufacture of retainers. Metallic coatings of lead, silver, nickel, cobalt, indium, and gold are applied.

6.12.9 Position of the Adjacent with Bearing Components: Drawing of the Interior Structure

In a bearing unit, the contact of the component adjacent with the bearing needs to be foreseen only along the faces of the bearing races, in the height of the collar. Other surfaces of the adjacent components must be at a distance from the faces of the races of not less than in 2–3 mm (the dimension a in Fig. 6.228) for all bearing classes (except tapered roller bearings).

The design of mating components with a radial double-row spherical bearing must take into account that, in some dimension types of these bearings, the balls jut out of the faces of the races by 0.7–2.8 mm.

A peculiarity of the structure of the tapered roller bearing is the fact that the retainer juts out of the ranges of the outer race by distances m and n, as shown in Fig. 6.229a. This should be taken into account by installation of the adjacent components with the bearing, e.g., of slotted nuts (Fig. 6.229b) or by mounting of two bearings positioned side by side (Fig. 6.229c). The adjacent component must be at a distance from the face of the outer race of the tapered roller bearing of $b = 4$–6 mm. So that the cylindrical surfaces of the adjacent components do not touch the retainer, the heights h_1 and h_2 must not exceed the values

$$h_1 = 0.1(D-d); \quad h_2 = 0.05(D-d)\,.$$

For bearings with a larger cone angle one uses $h_1 = 0.08(D-d)$.

Therefore, a spacer ring (1) is mounted between the faces of the inner bearing race and the nut in the most common fastening of the tapered bearing with a round slotted nut (Fig. 6.229b). Approximately half of the length of the ring (1) overlaps with the shaft of diameter d made for installation of the bearing, and the rest of the length overlaps with the groove for the outlet of the tool for thread cutting.

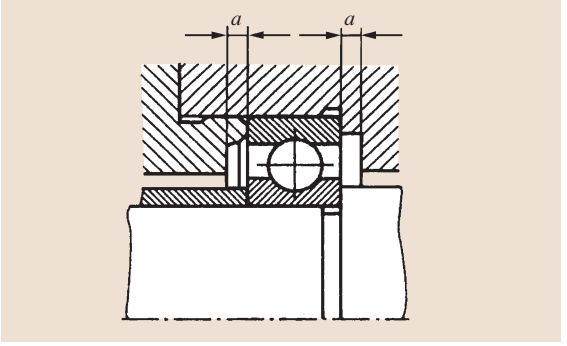


Fig. 6.228 Hookup of the adjacent components in the bearing unit

Drawing of the Interior Structure of Bearings

For the drawing of standard rolling bearings according to the overall dimensions d, D, and B the outer contour should be mapped with fine lines. Then for all bearing classes (except tapered roller bearings) the circle diameter $D_{pw} = 0.5(D+d)$ of the center position of the solids of revolution should be mapped. According to the formulas of Fig. 6.230a–e, the solids of revolution and rings are drawn.

Radial-thrust ball bearings (Fig. 6.230b) have only one ledge on the outer race. The second ledge is cut off. For sketching of the outer race from the side of the cut part an auxiliary vertical line is drawn to the intersection with the ball circle at point 1. The points 1 and 2

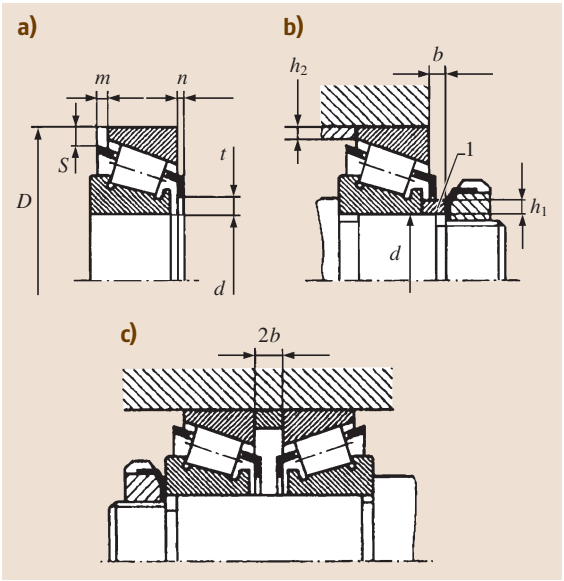


Fig. 6.229a–c Hookup of the adjacent components in the bearing unit with tapered roller bearings. **(a)** Peculiarities of the retainer structure, **(b)** allowable heights of the adjacent components, **(c)** allowable distance between two bearings

are joined. In ball radial double-row spherical bearings (Fig. 6.230c) the solids of revolution are represented in such a way that they do not touch the side lines of the outer contour. The spherical surface on the outer race is drawn as an arc of a circle with the center on the axis of the bearing hole.

For construction of tapered roller bearings (Fig. 6.230f) an auxiliary vertical line is drawn on the contour of the bearing that bisects the mounting height T of the bearing. Segment ab is divided by the points 1, 2, and 3 into four equal parts. From point 3 at an angle $\alpha = 15°$ the generating line of the cone is drawn up to its intersection with the pivot pin of the bearing at the point O. The lines O1 and O2 are drawn from this point. The segment $fk = 0.05(D - d)$ is mapped and the line fm is drawn transversely to the line O2. After the segment de is mapped, which is equal to fk, the line fm is drawn parallel to form a small face of the roller. To obtain the diameter d_2 of the ledge of the inner race a point

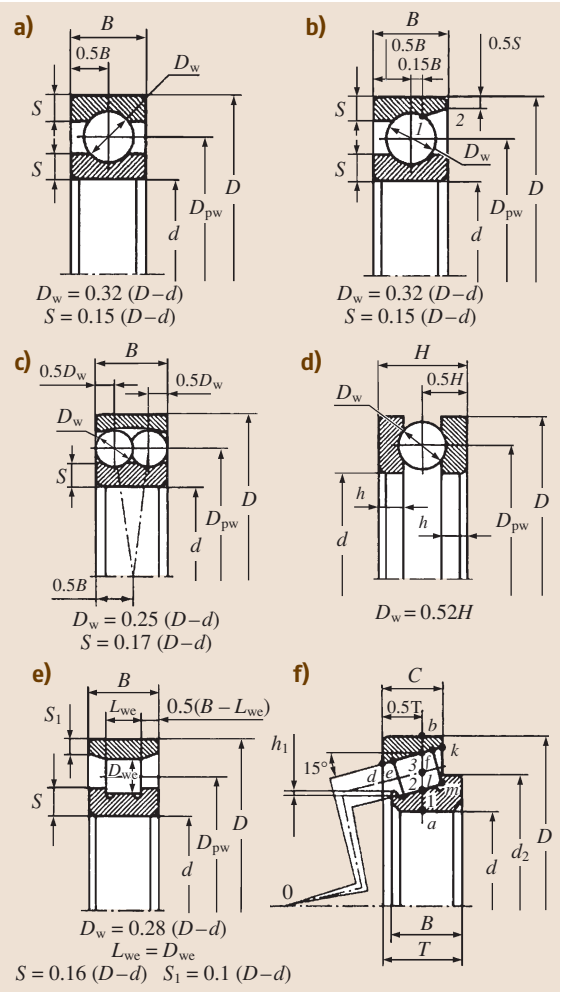


Fig. 6.230a–f Drawing of the inner structure of bearings. **(a)** Ball radial single-row bearing, **(b)** ball radial-thrust bearing, **(c)** ball radial double-row spherical bearing, **(d)** ball thrust bearing, **(e)** with short cylindrical rollers, **(f)** tapered roller bearing

l is found that bisects the radius of the bigger face of the roller. The height h_1 of a small ledge of the inner race is $h_1 = 0.124D_w$, where $D_w = fm$ is the largest diameter of the roller. The retainers in the drawings of the bearings are not depicted [6.116–118].

6.A Appendix A

Materials with similar chemical composition

Table 6.95 Steels

Grade	Country	Normative documentation
Ст5сп	Russia	ГОСТ
50 B	Great Britain	B.S.
St 50-2 G (E 295 + CR)	Germany	DIN
K 02305 // A 572 (50)	USA	ASTM
E 295 (A 50-2)	France	AFNOR NF; EN
SS 490	Japan	JIS
E 295	European standards	EN
Ст6сп	Russia	ГОСТ
55 C	Great Britain	B.S.
St 60-2 G (E 335 + CR)	Germany	DIN
A 572 (65)	USA	ASTM
E 335 (A 60-2)	France	AFNOR NF; EN
SM 570	Japan	JIS
E 355	European standards	EN
08кп	Russia	ГОСТ
1 HR	Great Britain	B.S.
2 HR	Great Britain	B.S.
3 HR	Great Britain	B.S.
DD 13 // StW 24	Germany	DIN; EN
DC 04 G1 // USt 4	Germany	DIN
A 622 (1008)	USA	ASTM
3 C	France	AFNOR NF
SPHE	Japan	JIS
10кп	Russia	ГОСТ
3 HR	Great Britain	B.S.
UStW 23 (DD 12 G1)	Germany	DIN
1010	USA	ASTM
2 C	France	AFNOR NF
SPHE70	Japan	JIS
35	Russia	ГОСТ
3	Great Britain	CEW
40 HS	Great Britain	B.S.
C 35	Great Britain	B.S.; EN
C 35 (C 35 k)	Germany	DIN; EN
G 10350 (1035)	USA	UNS
C 35	France	AFNOR NF
1572	Sweden	SS
S 35 C	Japan	JIS
C 36	European standards	EN
C 35	European standards	EN
45	Russia	ГОСТ
080 M 46	Great Britain	B.S.
C 45 E	Great Britain	B.S.; EN
C 45	Germany	DIN; EN
Cq 45	Germany	DIN
M 1044	USA	ASTM
C 45	France	AFNOR NF
1672	Sweden	SS
S 45 C	Japan	JIS
C 46	European standards	EN
C 45	European standards	EN
50	Russia	ГОСТ
060 A 52	Great Britain	B.S.
080 M 50	Great Britain	B.S.
C 50 E	Great Britain	B.S.
C 50 E // Ck 50	Germany	DIN; EN
G 10500 (1050)	USA	UNS
C 50 E (2 C 50)	France	AFNOR NF
1674	Sweden	SS
S 50 C	Japan	JIS
C 53	European standards	EN
C 50 E	European standards	EN

Grade	Country	Normative documentation
20Х	Russia	ГОСТ
207	Great Britain	B.S.
20 CrS 4	Germany	DIN
5120	USA	ASTM
SCr 420 H	Japan	JIS
40Х	Russia	ГОСТ
530 A 36	Great Britain	B.S.
530 M 40	Great Britain	B.S.
37 Cr 4	Germany	DIN; EN
41 Cr 4	Germany	DIN; EN
G 51400 (5140)	USA	UNS
H 51320	USA	UNS
H 51400 (5140 H)	USA	UNS
38 C 4	France	AFNOR NF
42 C 4	France	AFNOR NF
SCr 435	Japan	JIS
SCr 440	Japan	JIS
37 Cr 4	European standards	EN
41 Cr 4	European standards	EN
45Х	Russia	ГОСТ
5145	USA	ASTM
45 C 4	France	AFNOR NF
SCr 445 (SCr 5)	Japan	JIS
18ХГТ	Russia	ГОСТ
20 MnCr 5 G	Germany	DIN
30ХГТ	Russia	ГОСТ
30 MnCrTi	Germany	DIN
25ХГМ	Russia	ГОСТ
20 CrMo 5	Germany	DIN
35ХМ	Russia	ГОСТ
34 CrMo 4	Great Britain	B.S.; EN
34 CrMo	Germany	DIN; EN
G 41370 (4137)	USA	UNS
4135 H	USA	ASTM
34 CD 4	France	AFNOR NF
2234	Sweden	SS
SCM 435 H	Japan	JIS
34 CrMo 4 KD	European standards	EN
40ХН	Russia	ГОСТ
40 NiCr 6	Germany	DIN
3140 H	USA	AISI/SAE
G 31400 (3140)	USA	UNS
2530	Sweden	SS
SNC 236 (SNC 1)	Japan	JIS
12ХН3А	Russia	ГОСТ
14 NiCr 10 (5732)	Germany	DIN
10 NC 11	France	AFNOR NF
SNC 815 H	Japan	JIS
12Х2Н4А	Russia	ГОСТ
3310 H	USA	AISI/SAE
20ХН2М (20ХНМ)	Russia	ГОСТ
17 CrNiMo	Germany	DIN
4320	USA	AISI/SAE
SNCM 415	Japan	JIS
40ХН2МА (40ХНМА)	Russia	ГОСТ
36 CrNiMo 4 (817 M 37)	Great Britain	B.S.; EN
36 CrNiMo 4 (6511)	Germany	DIN; EN
G 43400 (4340)	USA	UNS
36 CrNiMo 4	France	AFNOR NF; EN
SNCM 439 (SNCM 8)	Japan	JIS
40 NiCrMo 4 KD	European standards	EN

Table 6.95 (cont.)

Grade	Country	Normative documentation
38X2MЮA (38ХМЮА)	Russia	ГОСТ
905 M 39	Great Britain	B.S.
41 CrAlMo 7	Germany	DIN
J 24056	USA	UNS
40 CAD 6.12	France	AFNOR NF
2940	Sweden	SS
SACM 645	Japan	JIS
41 CrAlMo 7	European standards	EN
65Г	Russia	ГОСТ
Ck 67	Germany	DIN
1566	USA	ASTM
55C2	Russia	ГОСТ
251 A 58	Great Britain	B.S.
55 Si 7	Germany	DIN
G 92550 (9255)	USA	UNS
55 S 7	France	AFNOR NF
2085	Sweden	SS
SUP 7	Japan	JIS
55 Si	European standards	EN
60C2A	Russia	ГОСТ
65 Si 7	Germany	DIN
9260	USA	ASTM
SUP 6	Japan	JIS
ШХ15	Russia	ГОСТ
100 Cr 6 (3505)	Germany	DIN; LW
J 19965 (52100)	USA	UNS
100 C 6	France	AFNOR NF
2258	Sweden	SS
SUJ 4	Japan	JIS
100 Cr 6	European standards	EN
ШХ15СГ	Russia	ГОСТ
100 CrMn 6 (3520)	Germany	DIN
K 19195	USA	UNS
100 CrMn 6 (100 CM 6)	France	AFNOR NF
100 CrMn 6	European standards	EN
12X18H10T	Russia	ГОСТ
321 S 31	Great Britain	B.S.
X 12 CrNiTi 18-9	Germany	DIN; SEW
Z10 CNT 18-10	France	AFNOR NF
2337	Sweden	SS
SUS 321	Japan	JIS
X 10 CrNiTi 18 10	European standards	EN
35Л	Russia	ГОСТ
GS-52	Germany	DIN
1	USA	ASTM
280-480 M (3)	France	AFNOR NF
SC 480	Japan	JIS
50Л	Russia	ГОСТ
GS-60	Germany	DIN
4A	USA	ASTM
370-650 M (3)	France	AFNOR NF
SCC5A	Japan	JIS

Table 6.96 Cast iron

Grade	Country	Normative documentation
ВЧ 50	Russia	ГОСТ
500/7	Great Britain	B.S.
GGG-50	Germany	DIN
70-50-05	USA	ASTM
FGS 500-7	France	AFNOR NF
FCD 500	Japan	JIS
500-7	ISO	ISO
ВЧ 60	Russia	ГОСТ
600/3	Great Britain	B.S.
GGG-60	Germany	DIN
80-55-06	USA	ASTM
FGS 600-3	France	AFNOR NF
FCD 600	Japan	JIS
600-3	ISO	ISO
СЧ 15	Russia	ГОСТ
150	Great Britain	B.S.
GG-15	Germany	DIN
25B	USA	ASTM
FC 150	Japan	JIS
150	ISO	ISO
СЧ 20	Russia	ГОСТ
200	Great Britain	B.S.
GG-20	Germany	DIN
30B	USA	ASTM
FC 200	Japan	JIS
200	ISO	ISO
СЧ 25	Russia	ГОСТ
250	Great Britain	B.S.
GG-25	Germany	DIN
35B	USA	ASTM
FC 250	Japan	JIS
250	ISO	ISO
СЧ 30	Russia	ГОСТ
300	Great Britain	B.S.
GG-30	Germany	DIN
45B	USA	ASTM
FC 300	Japan	JIS
300	ISO	ISO
СЧ 35	Russia	ГОСТ
350	Great Britain	B.S.
GG-35	Germany	DIN
50B	USA	ASTM
FC 350	Japan	JIS
350	ISO	ISO

Table 6.97 Bronzes

Grade	Country	Normative documentation
БрО5Ц5С5	Russia	ГОСТ
C83800	USA	ASTM
C93500	USA	ASTM
C93200	USA	ASTM
Rg 5 (2.1097)	Germany	DIN
Rg 7 (2.1091)	Germany	DIN
H5111/class 6,6C	Japan	JIS
БрО5Ц6С5	Russia	ГОСТ
GB-CuSn5ZnPb	Germany	DIN
BC1.6	Japan	JIS
БрО10Ф1	Russia	ГОСТ
C90700	USA	ASTM
C90800	USA	ASTM
C91100	USA	ASTM
C91300	USA	ASTM
H5113/class 2	Japan	JIS
H5113/class 2B	Japan	JIS
H5113/class 2C	Japan	JIS
БрА9Ж3Л	Russia	ГОСТ
C95200	USA	ASTM
C95400	USA	ASTM
C95900	USA	ASTM
FeAlBz (2.0941)	Germany	DIN
H5114/class 1	Japan	JIS
БрА10Ж4Н4Л	Russia	ГОСТ
NiAlBz (2.0971)	Germany	DIN
H5114/class 3	Japan	JIS

6.B Appendix B

Conformity of accuracy grades of the bearings according to [6.106] with accuracy grades according to international standards and national standards of some countries.

Table 6.98 Ball and roller radial and radial-thrust ball bearings

National standard		Accuracy degree							
Interstate standard of CIS	ГОСТ 520	8	7	Normal	6	5	4	T	2
International organization for standardization	ISO 492	-	-	Normal	6	5	4	-	2
Standard of Germany	DIN 620	-	-	PO	P6	P5	P4	-	P2
Standard of the USA	AFBMA Standard 20	-	-	ABEC-1 RBEC-1	ABEC-3 RBEC-3	ABEC-5 RBEC-5	ABEC-7		ABEC-9
Standard of Japan	JISB B 1514	-	-	0	6	5	4	-	2

Table 6.99 Roller tapered bearings

National standard		Accuracy degree								
Interstate standard of CIS	ГОСТ 520	8	7	0	Normal	6X	6	5	4	2
International organization for standardization	ISO 492	-	-	-	Normal	6X	-	5	4	2
Standard of Germany	DIN 620	-	-	-	PO	P6X	-	P5	P4	P2
Standard of the USA	AFBMA Standard 19.1	-	-	-	K	N	-	C	B	A
Standard of Japan	JISB B 1514	-	-	-	0	6X	6	5	4	2

Table 6.100 Thrust and thrust-radial bearings

National standard		Accuracy degree						
Interstate standard of CIS	ГОСТ 520	8	7	Normal	6	5	4	2
International organization for standardization	ISO 492	-	-	Normal	6	5	4	-
Standard of Germany	DIN 620	-	-	PO	P6	P5	P4	-

References

6.1 L.A. Andrienko, B.A. Baykov, I.K. Ganulich: *Components of Machines*, ed. by O.A. Riakhovsky (Moscow State Technical Bauman University, Moscow 2004), in Russian

6.2 V.I. Anuriev: *Handbook of an Engineer-Technician*, Vol. 1–3, 9th edn. (Mashinostroenie, Moscow 2006), in Russian

6.3 I.A. Birger, B.F. Shorr, G.B. Iosilevich: *Calculation of the Strength of Maschines Components* (Mashinostroenie, Moscow 1993), in Russian

6.4 K.H. Decker: *Maschinenelemente – Funktion, Gestaltung und Berechnung*, 16th edn. (Hanser, München 2007), in German

6.5 K.H. Grote, J. Feldhusen (Eds.): *Dubbel – Taschenbuch für den Maschinenbau*, 22nd edn. (Springer, Berlin Heidelberg 2005), in German

6.6 K.V. Frolov, A.F. Kraynev, G.V. Kreynin: *Design of Machines*, Vol. 2 (Mashinostroenie, Moscow 1994), in Russian

6.7 V.I. Feodosiev: *Resistance of Materials*, 10th edn. (Publishing House of Moscow State Technical Bauman University, Moscow 2000), in Russian

6.8 H. Czichos, M. Hennecke (Eds.): *Hütte – Das Ingenieurwissen*, 33rd edn. (Springer, Berlin Heidelberg 1996), in German

6.9 M.N. Ivanov, V.A. Finogenov: *Components of Machines*, 9th edn. (Vysshaya Shkola, Moscow 2005), in Russian

6.10 V.V. Klyuev (Ed.): *Mechanical Engineering. Encyclopedia. Safety of Machines*, Vol. IV-3 (Mashinostroenie, Moscow 1998), in Russian

6.11 O.P. Lelikov: *Calculation and Design of the Components and Units*, 3rd edn. (Mashinostroenie, Moscow 2007), in Russian

6.12 G. Niemann, H. Winter: *Components of Machines*, Vol. II and III, 2nd edn. (Springer, Berlin Heidelberg 1983), in German

6.13 D.N. Reshetov (Ed.): *Encyclopedia of Mechanical Engineering. Machine Parts, Friction, Wear, Lubrication*, Vol. IV-1 (Mashinostroenie, Moscow 1995), in Russion

6.14 V.M. Trukhanov: *Safety in Machinery* (Mashinostroenie, Moscow 1999)

6.15 R.J. Drago: *Fundamentals of Gear Design* (Butterworth, Boston 1988)

6.16 S.P.A. Bonfiglioli Riduttori, D.W. Dudley, J. Sprengers, D. Schröder, H. Yamashina: *Gear Motor Handbook* (Springer, Berlin Heidelberg 1995)

6.17 D.W. Dudley: *Handbook of Practical Gear Design* (CRC, Boca Raton 1994)

6.18 K. Johnson: *Mechanics of contact interaction* (Mir, Moscow 1989), translated from English

6.19 V.P. Kogaev: *Strength analysis by stresses variable in time*, 2nd edn., ed. by A.P. Gusenkov (Mashinostroenie, Moscow 1993), in Russian

6.20 V.N. Kudriavtsev, I.S. Kuzmin, A.A. Filipenkov: *Calculation and Design of Reduction Gear Units*, ed. by V.N Kudriavtsev (Politechnika, St. Petersburg 1993), in Russian

6.21 A.V. Orlov, O.N. Chermensky, V.M. Nesterov: *Contact Fatigue Testing of Structural Materials* (Mashinostroenie, Moscow 1980), in Russian

6.22 G. Pahl, W. Beitz, J. Feldhusen, K.-H. Grote: *Konstruktionslehre*, 3rd edn. (Springer, London 1997)
6.23 I.E. Shigley: *Mechanical Engineering Design* (McGraw-Hill, New York 1977)
6.24 E.B. Vulgakov (Ed.): *Aviation Gearings and Reduction Gears* (Mashinostroenie, Moscow 1981), in Russian
6.25 E.B. Vulgakov: *Coaxial Gearings* (Mashinostroenie, Moscow 1987), in Russian
6.26 V.P. Kogaev, I.V. Gadolina: Summation of fatigue damages by probability calculation of service life, Vestn. Mashinostr. **7**, 3–7 (1989), in Russian
6.27 GOST 25.587-78 Calculations and strength tests in mechanical engineering. Test methods of contact fatigue (Standards Publishing House, Moscow 1978)
6.28 GOST 1643-81 Principal standards of interchangeability. Cylindrical gearings. Tolerances (Standards Publishing House, Moscow 1981)
6.29 GOST 1758-81 Bevel and hypoid gears. Tolerances (Standards Publishing House, Moscow 1981)
6.30 GOST 9563-60 Principal standards of interchangeability. Gear wheels. Modules (Standards Publishing House, Moscow 1960)
6.31 GOST 13754-81 Principal standards of interchangeability. Bevel gearings with straight teet. Original profile (Standards Publishing House, Moscow 1981)
6.32 GOST 13755-81 Principal standards of interchangeability. Involute gears. Original profile (Standards Publishing House, Moscow 1981)
6.33 GOST 19326-73 Bevel gearings with circular teeth. Calculation of geometry (Standards Publishing House, Moscow 1973)
6.34 GOST 19624-74 Bevel gearings with straight teeth. Calculation of geometry (Standards Publishing House, Moscow 1974)
6.35 GOST 21354-87 Cylindrical involute gearings of external toothing. Strength analysis (Standards Publishing House, Moscow 1987)
6.36 GOST R 50891-96 Reduction gears of machine-building application. General technical conditions (Standards Publishing House, Moscow 1996)
6.37 GOST R 50968-96 Reduction gearmotors. General technical conditions (Standards Publishing House, Moscow 1996)
6.38 E.L. Airapetov, M.D. Genkin, T.N. Melnikova: *Static of Globoidal Gears* (Nauka, Moscow 1981), in Russian
6.39 G. Niemann, H. Winter: *Machinenelemente*, 2nd edn. (Springer, Berlin Heidelberg 1983), in German
6.40 V.V. Shults: *Natural Wear-and-Tear of Machine Components and Tools* (Mashinostroenie, Leningrad 1990), in Russian
6.41 GOST 3675-81 Principal standards of interchangeability. Worm cylindrical gearings. Tolerances (Standards Publishing House, Moscow 1981)
6.42 GOST 16502-83 Principal standards of interchangeability. Globoidal gears. Tolerances (Standards Publishing House, Moscow 1983)
6.43 GOST 17696-89 Globoidal gears. Calculation of geometry (Standards Publishing House, Moscow 1989)
6.44 GOST 19036-94 Principal standards of interchangeability. Worm cylindrical gearings. Original worm and original productive worm (Standards Publishing House, Moscow 1994)
6.45 GOST 19650-97 Worm cylindrical gearings. Calculation of geometry (Standards Publishing House, Moscow 1997)
6.46 GOST 19672-74 Worm cylindrical gearings. Modules and coefficients of the worm diameter (Standards Publishing House, Moscow 1974)
6.47 GOST 24438-80 Globoidal gears. Original worm and original productive worm (Standards Publishing House, Moscow 1980)
6.48 P.F. Dunaev, O.P. Lelikov: *Design of Units and Components of Machines*, 9th edn. (Academy, Moscow 2006), in Russian
6.49 P.F. Dunaev, O.P. Lelikov: *Calculation of Dimensional Tolerances*, 4th edn. (Mashinostroenie, Moscow 2006), in Russian
6.50 P.F. Dunaev, O.P. Lelikov: *Components of machines*, 5th edn. (Mashinostroenie, Moscow 2007), in Russian
6.51 P.F. Dunaev, O.P. Lelikov, L.P. Varlamova: *Tolerances and Fits* (Vysshaya shkola, Moscow 1984), in Russian
6.52 GOST 2.309-73 (edn. 2003) Uniform system of design documentation. Designations of surface roughness and marking regulations in the drawings of products (Standards Publishing House, Moscow 2003)
6.53 GOST 25346-89 Uniform system of tolerances and fits. General provisions, series of tolerances and principal deviations (Standards Publishing House, Moscow 1989)
6.54 GOST 25347-82 Uniform system of tolerances and fits. Tolerance ranges and advisable fits (Standards Publishing House, Moscow 1982)
6.55 GOST 30893.1-2002 (ISO 2768-1-89) Principal standards of interchangeability. General tolerances. Extreme deviations of linear and angular dimensions with non-specified tolerances (Standards Publishing House, Moscow 2002)
6.56 GOST 30893.2-2002 (ISO 2768-2-89) Principal standards of interchangeability. General tolerances. Tolerances of form and position of surfaces non-specified individually (Standards Publishing House, Moscow 2002)
6.57 E.L. Ayrapetov, M.D. Genkin: *Dynamics of Planetary Trains* (Nauka, Moscow 1980), in Russian
6.58 E.G. Ginzburg: *Wave Gears* (Mashinostroenie, Leningrad 1969), in Russian
6.59 M.N. Ivanov: *Wave Gears* (Vysshaya Shkola, Moscow 1981), in Russian
6.60 V.N. Kudriavtsev: *Planetary Gears* (Mashinostroenie, Moscow, Leningrad 1966), in Russian
6.61 V.N. Kudriavtsev, Y.N. Kirdiashev (Eds.): *Planetary Gears: Reference Book* (Mashinostroenie, Moscow 1977), in Russian

6.62 GOST 9587-81 Principal standards of interchangeability. Gearings Original profile of fine-module gear wheels (Standards Publishing House, Moscow 1981)

6.63 GOST 10059-80 Fine-module finishing gear-shaping cutters. Technical conditions (Standards Publishing House, Moscow 1980)

6.64 GOST 23179-78 Radial ball single-row flexible rolling bearings. Technical conditions (Standards Publishing House, Moscow 1978)

6.65 GOST 25022-81 Planetary gearboxes. Critical parameters (Standards Publishing House, Moscow 1981)

6.66 GOST 26218-94 Harmonic reduction gears and reduction gearmotors. Parameters and dimensions (Standards Publishing House, Moscow 1994)

6.67 GOST 26543-94 Planetary reduction gearmotors. Critical parameters (Standards Publishing House, Moscow 1994)

6.68 GOST 30078.1-93 Wave gears. General technical requirements (Standards Publishing House, Moscow 1993)

6.69 GOST 30078.2-93 Wave gears. Types. Critical parameters and dimensions (Standards Publishing House, Moscow 1993)

6.70 V.L. Biderman: *Theory of Mechanical Oscillations* (Vysshaya Shkola, Moscow 1980), in Russian

6.71 V.V. Bolotin: *Vibration in Engineering: Handbook*, Vol. 1–6 (Mashinostroenie, Moscow 1978), in Russian

6.72 O.P. Lelikov: *Shafts and Supports with Frictionless Bearings* (Mashinostroenie, Moscow 2006), in Russian

6.73 G.S. Maslov: *Calculation of the Vibration of Shafts*, 2nd edn. (Mashinostroenie, Moscow 1980), in Russian

6.74 S.V. Serensen, M.B. Groman, V.P. Kogaev, R.M. Shneiderovich: *Shafts and Axles* (Mashinostroenie, Moscow 1970), in Russian

6.75 S.V. Serensen, V.P. Kogaev, R.M. Shneiderovich: *Load-Carrying Ability and Strength Analysis of the Machine Components*, 3rd edn. (Mashinostroenie, Moscow 1975), in Russian

6.76 W. Steinhilper, R. Röper: *Maschinenelemente*, Vol. 1-3, 4th edn. (Springer, Berlin Heidelberg 1994), in German

6.77 W. Weaver, S.P. Timoshenko, D.H. Young: *Vibration Problems in Engineering* (Wiley Interscience, New York 1985)

6.78 R50-83-88 Recommendations: *Calculations and Strength Testing. Strength Analysis of Shafts and Axles* (Publishing House of Standards, Moscow 1989)

6.79 GOST 2.307-68 Uniform system of design documentation. Marking of dimensions and extreme deviations (Standards Publishing House, Moscow 1968)

6.80 GOST 2.308-79 Uniform system of design documentation. Indication of form and surface position tolerances in the drawings (Standards Publishing House, Moscow 1979)

6.81 GOST 25.504-82 Calculations and strength testing. Calculation methods of fatigue strength characteristics (Standards Publishing House, Moscow 1982)

6.82 GOST 2789-73 Surface roughness. Parameters and characteristics (Standards Publishing House, Moscow 1973)

6.83 GOST 6636-69 Normal linear dimensions (Standards Publishing House, Moscow 1969)

6.84 GOST 12080-66 Cylindrical shaft ends. Basic dimensions, allowable torsional moments (Standards Publishing House, Moscow 1966)

6.85 GOST 12081-72 Tapered shaft ends with a taper 1:10. Basic dimensions, allowable torsional moments (Standards Publishing House, Moscow 1972)

6.86 GOST 22061-76 Machines and processing equipment. System of balancing accuracy classes (Standards Publishing House, Moscow 1976)

6.87 GOST 24266-94 Shaft ends of reduction gears and reduction gearmotors. Basic dimensions, allowable torsional moments (Standards Publishing House, Moscow 1994)

6.88 GOST 24643-81 Principal standards of interchangeability. Tolerances of form and surface positions. Values (Standards Publishing House, Moscow 1981)

6.89 GOST 3325-85 Rolling bearings. Tolerance ranges and technical requirements for mounting surfaces of the shafts and cases. Fits (Standards Publishing House, Moscow 1985)

6.90 GOST 23360-78 Feather keys. Dimensions, tolerances and fits (Standards Publishing House, Moscow 1978)

6.91 G.A. Bobrovnikov: *Strength of Force Fits Attained by Cooling* (Mashinostroenie, Moscow, 1971), in Russian

6.92 E.S. Grechishchev, A.A. Il'iashenko: *Joints with Interference* (Mashinostroenie, Moscow 1981), in Russian

6.93 K. Ootsuka, K. Simidzu, Y. Sudzuki: *Alloys with an Effect of Shape Memory*, ed. by H. Funakubo (Metallurgia, Moscow 1990), in Russian

6.94 A.A. Illin: Alloys with an effect of shape memory. Totals of science and technology, Phys. Met. Heat Treat. **25**, 3–39 (1991)

6.95 D.N. Reshetov, Y.V. Krasnov: Statistical analysis of friction coefficient in the joints with interference, Isvestia Vuzov. Mashinostr. **4**, 15–19 (1985), in Russian

6.96 GOST 1139-80 Principal standards of interchangeability. Straight-sided spline connections. Dimensions and tolerances (Standards Publishing House, Moscow 1980)

6.97 GOST 6033-80 Principal standards of interchangeability. Involute spline connections with a profile angle 30°. Dimensions, tolerances and measurable values (Standards Publishing House, Moscow 1980)

6.98 GOST 21425-75 Straight-sided serrated (spline) joints. Calculation methods of load-carrying ability (Standards Publishing House, Moscow 1975)

6.99 GOST 24071-80 Principal standards of interchangeability. Key joints with semicircular keys. Dimen-

sions of keys and groove sections. Tolerances and fits (Standards Publishing House, Moscow 1980)
6.100 J. Brändlein, P. Eschmann, L. Hasbargen, K. Weigang: *Wälzlagerpraxis* (Vereinigte Fachverlage GmbH, Mainz 1995), in German
6.101 P. Eschmann, L. Hasbargen, K. Weigang: *Ball and Roller Bearings* (Wiley, New York 1985)
6.102 T.A. Harris: *Rolling Bearing Analysis*, 4th edn. (Wiley, New York 2000)
6.103 L.Y. Perel, A.A. Filatov: *Rolling Bearing* (Mashinostroenie, Moscow 1992), in Russian
6.104 SKF Cataloque 6000EN, November 2005 (SKF, Schweinfurt 2005)
6.105 D.N. Reshetov, O.P. Lelikov: Calculation of rolling bearings by varying loads, Isvestia Vuzov Mashinostr. **12**, 15–19 (1984)
6.106 GOST 520-2002 Rolling bearings. General technical conditions (Standards Publishing House, Moscow 2002)
6.107 GOST 3189-89 Ball and roller bearings. Nomenclature (Standards Publishing House, Moscow 1989)
6.108 GOST 3395-89 Rolling bearings. Classes and embodiments (Standards Publishing House, Moscow 1989)
6.109 GOST 13942-86 Spring thrust planar outer eccentric rings and grooves for them. Structure and dimensions (Standards Publishing House, Moscow 1986)
6.110 GOST 13943-86 Spring thrust planar inner eccentric rings and grooves for them. Structure and dimensions (Standards Publishing House, Moscow 1986)
6.111 GOST 18854-94 (ISO 76-87) Rolling bearings. Static load rating (Standards Publishing House, Moscow 1994)
6.112 GOST 18855-94 (ISO 281-89) Rolling bearings. Dynamic rated load rating and design life (Standards Publishing House, Moscow 1994)
6.113 GOST 20226-82 Collars for installation of rolling bearings. Dimensions (Standards Publishing House, Moscow 1982)
6.114 GOST 24810-81 Rolling bearings. Clearances. Dimensions (Standards Publishing House, Moscow 1981)
6.115 ISO 5593-84 Rolling bearings. Terminological dictionary
6.116 E.A. Chernyshov: *Casting Alloys and their Foreign Analogs* (Mashinostroenie, Moscow 2006), in Russian
6.117 O.E. Osintsev, V.N. Fedotov: *Copper and Copper Alloys. Russian and Foreign Brands* (Mashinostroenie, Moscow 2004), in Russian
6.118 A.S. Zubchenko (Ed.): *Grades of Steels and Alloys*, 2nd edn. (Mashinostroenie, Moscow 2003), in Russian

7. Manufacturing Engineering

Thomas Böllinghaus, Gerry Byrne, Boris Ilich Cherpakov (deceased), Edward Chlebus, Carl E. Cross, Berend Denkena, Ulrich Dilthey, Takeshi Hatsuzawa, Klaus Herfurth, Horst Herold (deceased), Andrew Kaldos, Thomas Kannengiesser, Michail Karpenko, Bernhard Karpuschewski, Manuel Marya, Surendar K. Marya, Klaus-Jürgen Matthes, Klaus Middeldorf, Joao Fernando G. Oliveira, Jörg Pieschel, Didier M. Priem, Frank Riedel, Markus Schleser, A. Erman Tekkaya, Marcel Todtermuschke, Anatole Vereschaka, Detlef von Hofe, Nikolaus Wagner, Johannes Wodara, Klaus Woeste

Manufacturing is the set of activities converting raw materials into products in the most possible cost effective way, including design of goods, manufacturing parts and assembling them into products (subassemblies) using various production methods and techniques, the sale of products to customers, servicing, maintaining the product in good working order, and eventually recycling materials and parts. Whilst the design stage costs about 10–15% of all manufacturing costs, its effect on all other activities is enormous. The designed product has to be easy to make, easy to assemble, maintainable at a competitive cost level, and finally it should be economically recyclable. This is why concurrent engineering (CE) is a systematic approach integrating the design stage and manufacturing stage of products with a view to optimizing all elements involved in the life cycle of a product.

Due to the vast complexity of manufacturing engineering it can only be dealt with in a number of different chapters. The sections in this chapter illustrate the most important manufacturing processes from casting to assembly, from the first shape giving process to the last component integrative process. In between the reader will find a variety of manufacturing processes, including the most recent technologies, e.g. microbonding, nanotechnology, and others. Chapter 10 describes the front end of manufacturing, i. e. design, and Chap. 16 is allocated to quality assurance in manufacturing engineering. Finally, Chap. 17 is devoted to manufacturing logistics and manufacturing system analysis.

Manufacturing creates new wealth by its various value adding activities, including the addition of human knowledge and expertise to the products. Manufacturing provides products, which themselves are used to make other products and services, e.g. the machine tool industry is of vital importance for the well-being of a nation.

Manufacturing is the backbone of any industrialized nation contributing with approximately 20–30% of the value of all goods and services produced. However, it is increasingly difficult to meet customer demands and compete in the global market. Therefore manufacturing must be able to react speedily to changing market demands and to maximize the utilization of all resources, e.g. human (man), financial (money), equipment and buildings (machine), and time.

Manufacturing needs information on component parts to be produced (material, shape, features, size, accuracy/finish requirements, and batch size. This determines/affects the selection of most appropriate manufacturing technologies to be used, together with the selection of raw material, billets, etc., tool material (insert), tool life, tool assembly, tool type to be used together with the selection of type of machine tools (M/T) to be used, capabilities of M/T, work envelope, accuracy, power available, type of control, make of control, automatic tool change (ATC), pallet/workpiece changing facilities, swarf management, chip removal, tool interface type, tool holders, adapters required, accuracy, range of operational parameters (spindle speeds, feed rates, etc.), type of drive system, tool storage facilities, and others. This leads to process planning and manufacturing planning for the manufacturing sites. It is also very important to decide on part location and fixing and select or design the most suitable jigs, fixtures, and clamping.

Information related to all manufacturing activities is compiled in an integrated database with relevant access to any level of activity. However, it is of vital importance to verify data input and remove outdated information.

As manufacturing relies heavily on human involvement even in the case of fully automated systems, human ethics must not be violated or even overstretched due to the anatomy of the human, and thus requirements of ergonomics have to be met.

Not only must products fully meet design requirements and specifications, they have to be manufactured by the most economical methods in order to minimize costs.

Furthermore, quality must be manufactured into the product at each stage; the final testing is not enough. Therefore quality assurance (QA) plays a vital role in manufacturing from design through parts manufacture and assembly to after sales services. In a highly competitive market manufacturing methods must be sufficiently flexible to respond to changing market demands, types of products, production rates, production quantities, and on-time deliveries.

New developments in materials, production methods and computer integration of both technological and managerial activities in a manufacturing organization must constantly be evaluated with the view of imple-

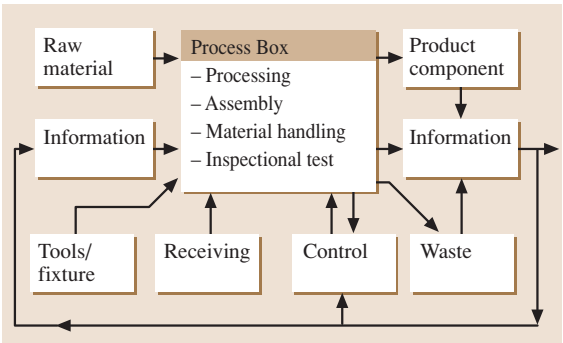


Fig. 7.1 A simplified process model for parts/product manufacturing

mentation. New materials meet new challenges at the design and parts manufacturing stages alike.

Manufacturing must be viewed as a system of activities and subsystems that are interrelated to others. These systems can be modeled to study the effect of various input factors.

A simplified process model can be used to illustrate parts/product manufacturing with input and output parameters as shown in Fig. 7.1.

It is important to note the role of information in the system and the number of feedback loops required to maintain the various functions in manufacturing.

7.1 Casting

7.1.1 The Manufacturing Process

Manufacturing is the production of workpieces of a geometrically defined shape. Unlike other production technologies, manufacturing technology produces products distinguished by material and geometric characteristics.

According to DIN 8580 [7.1] today's and tomorrow's many and varied manufacturing processes can be classified into six main groups (Fig. 7.2) according to the criteria *changing of material cohesion* (creation, preservation, increase, and reduction). These main groups are: primary shaping, forming, cutting, joining, coating, and changing of material properties. The main groups are subdivided into groups. Within these groups, the manufacturing processes themselves are distinguished by subgroups.

Casting is an important group of the main group *primary shaping*. The products of casting are so-called castings from metallic materials, so-called cast alloys. The foundry industry produces the castings.

7.1.2 The Foundry Industry

The foundry industry produces castings from metallic materials. Castings in various shapes and materials are to be found in all branches of engineering, such as in road and rail vehicles, machinery, the aerospace industry, electric power engineering, shipbuilding, pumps and fittings, electrical engineering, fine mechanics, architecture, electronics, medicine and optics, as well as office machines and cast art. Most technical products are not conceivable without castings.

Create cohesion	Preserve cohesion	Reduce cohesion	Increase cohesion	
	Shape changing			
1. Primary shaping	2. Metal forming	3. Cutting	4. Joining	5. Coating
	6. Changing of material properties			
	Rearrangement of material particles	Elimination of material particles	Addition of material particles	

Fig. 7.2 Classification of manufacturing processes (DIN 8580)

Advantages of Casting Production

The following outlines the economical advantages resulting from the use of castings. It is based on two fundamental aspects: firstly, the overall consideration from melting to recycling, namely, recirculation of the metals within the framework of the national economy, and secondly, a comparison with other main groups of manufacturing such as forming, cutting and joining in accordance with DIN 8580 [7.1]. The advantages are the following:

- Exceptional freedom in the choice of shape
- Castability of all technically important metals
- Mechanical properties of cast materials that are no longer governed by those of the wrought materials
- Optimal components through the combination of material properties and shape
- Near net shape castings, i. e. reduced machining costs
- Integral castings, i. e. reduced assembly costs
- Tailored castings
- A high recycling rate, no down-cycling
- Material savings through the use of castings
- Ecological benefits
- When considered in overall terms, a lower specific use of energy on the whole

Production from the shapeless condition, i. e. the melt, and solidification in the prescribed cavity, i. e. sand or permanent mold, naturally enables great freedom in the choice of shape. It is possible to produce castings weighing less than 1 g and as much as 200 t, the weight only being restricted by handling and transport facilities.

There are very few restrictions with regard to the geometry of castings. All that can be drawn is also castable, but it is sometimes difficult to draw what is easy to cast.

High quality casting designs increasingly result from the incorporation of numerical simulation of mold filling and solidification, rapid prototyping, and simultaneous engineering. This takes place through close cooperation between foundrymen and designers. In the future it is expected that bionic and biological designing will provide new impulses for shaping. This will enable a large amount of freedom in the choice, thereby leading to full utilization of modern computer technology in the foundry industry.

Properties of Castings

Castings are produced from the following material groups: iron alloys (cast steel, cast iron), aluminum, magnesium, titanium, copper, zinc, tin, nickel, and cobalt alloys. All of these are cast alloys.

Independent of the type of production process, in the manufacture of metallic components by casting differentiation is always to be made between the material properties and the properties of the casting itself. In order to achieve a prescribed component characteristic the material and the geometry determine and complement each other in the properties of the component. These properties depend on the following:

- The geometry of the part
- The chemical composition of the cast material
- The treatment of the melted material (inoculation, modification, desulfurization, deoxidation, magnesium treatment, grain refinement, etc.)
- The type of molding and casting process
- The rate of cooling from casting to the ambient temperature
- The subsequent heat treatment
- The subsequent treatment of the outer layer (chemical-thermal process, surface deformation, surface alloying, surface remelting, etc.)
- Changes in the surface layer through machining
- The type of coating (painting, galvanizing, enameling etc.)

During the past decades the properties of cast alloys have also been further developed and considerably improved. For example, whereas in the 1950s only steel casting was able to achieve tensile strengths of more than 400 N/mm^2, today the designer has the choice of three higher strength groups of ferrous materials, i. e. spheroidal graphite cast iron, malleable cast iron, and cast steels. In many cases, this has enabled the highly economical substitution of forged and rolled steels.

There have also been further developments in nonferrous metal cast alloys, especially in the fields of aluminum and magnesium materials, which increasingly enable the use of these alloys in automotive manufacturing.

Consequently, the last decade has seen substantial rates of growth in the production of spheroidal graphite cast iron, as well as in aluminum and magnesium alloy castings. This is directly associated with the efforts towards light construction but also towards the reduction of total production costs.

The trend towards light construction is not only being realized with less dense cast materials, e.g. aluminum, but is also being achieved with higher density materials, e.g. spheroidal graphite cast iron. This is the result of the combined effect of material and shape as well as further development of casting technology.

In the selling of castings the material properties were traditionally (and often still are) taken as a basis for the contract, i. e. such material characteristics as yield strength, tensile strength, elongation at fracture, fatigue strength etc., which were determined from separately quasi simple geometrical samples. However, these samples only partially reflect the capabilities of the cast materials. Cast components are increasingly being designed on the basis of fracture mechanics. This shows that cast components are frequently unbeatable.

The Development of Casting Processes

The development of molding and casting processes during recent decades has led to the fact that the casting more and more approaches the shape of the finished part. The best results have been achieved with investment casting (lost wax molding process) and high pressure die casting, which can produce almost finished parts. These often require only minimum machining, e.g. fine machining of operating surfaces. Additionally, the development of weldable aluminum pressure die castings has enabled further possibilities of use.

Mechanical machining requires a relatively high amount of energy, the generation of 1 t of chips requiring the same amount of energy as that for the melting of 1 t of material.

The chips produced in machining the casting to the complete product, which can often result in a material utilization of less than 50%, should now be a thing of the past. The future lies in the production of near net shape castings with the resultant large savings in energy.

In many cases, groups of parts were and are being assembled from numerous individual components (turned, milled, and sheet metal components) by means of welding, riveting, bolting, etc. This type of assembly not only necessitates expensive manufacture of individual parts but also gives rise to considerable assembly costs.

The casting of integral components (one-piece-castings), through which the numerous previously necessary individual parts are combined in one casting, is an ideal way towards a new generation of parts. These integral castings can additionally better incorporate specific functional elements, resulting in considerable savings in material and energy.

Recycling is understood to be the return of material into the production process. In doing so, the aim is not to leave industrial production open-ended but, as with nature, to close the circuit, here material flows.

Recycling is in no way a new term for metallic materials but is rather a thousands of years old practice of returning metallic waste into the production process. Recycling of cast steel, cast iron, and cast nonferrous metals is a worldwide normal practice. Recycling of metallic materials leads to the saving of energy, preservation of our raw materials reserves, and thus to relief of our environment.

Development of the properties of the cast materials and the improvement of the molding and casting processes in the foundry industry have not only led to higher productivity but also material saving through:

- The reduction of wall thicknesses as a result of better pouring possibilities
- The use of higher strength materials
- Optimal casting design with, for example, ribbing or realization of hollow structures
- The reduction of machining allowances
- Material substitution, e.g. spheroidal graphite cast iron instead of forged steel and aluminum alloys instead of cast iron

These savings in materials reduce the weight of the components as well as the amount of machining. They also result in energy savings and thus preservation of the environment.

Overall consideration of component manufacture from the raw material to the finished part and taking account of recycling of metallic materials, i. e. an economic balance, illustrated that, by comparison with the other main production processes, the manufacture and use of castings results in substantial energy savings and thus ecological advantages, e.g. the reduction of CO_2.

From case studies, it is a well-known fact that, by comparison with other process variants, near net shape castings are clearly advantageous with regard to specific energy requirements for the finished components, especially with respect to machining from solid semiproducts. Component manufacture by casting is also clearly preferential when considering ecological aspects such as CO_2 emission.

Foundries pursue objective environmental management and confront public option with declarations of their achievements.

7.1.3 Cast Alloys

Cast alloys are metallic materials manufactured by primary shaping in a foundry. Cast alloys can be classified in two main groups: cast ferrous materials (cast irons and cast steels) and cast nonferrous materials (cast aluminum, cast magnesium, cast copper, and cast zinc alloys).

Cast Iron Alloys

Cast iron alloys can be classified into seven groups:

- Gray cast iron
- Spheroidal graphite cast iron
- Ausferrite spheroidal graphite cast iron
- Compacted graphite cast iron
- Malleable cast iron
- Austenitic cast iron
- Abrasion resisting alloyed cast iron

Cast Iron. The term *cast iron* designates an entire family of metallic materials with a wide variety of properties. It is a generic term like steel, which also designates a family of metallic materials. Steels and cast irons are both primarily iron with carbon as the main alloying element. Steels contain less than 2%, while all cast irons contain more than 2% carbon. About 2% is the maximum carbon content at which iron can solidify as a single phase alloy with all of the carbon in solution in austenite. Thus, the cast irons by definition solidify as heterogeneous alloys and always have more than one constituent in their microstructure.

In addition to carbon, cast irons also must contain appreciable silicon, usually from 1 to 3%, and thus they are actually iron-carbon-silicon alloys. The high carbon content and the silicon in cast irons make them excellent casting alloys. Their melting temperatures are appreciably lower than those of steel. Molten cast irons are more fluid than molten steel and less reactive with molding

materials. Formation of lower density graphite in cast iron during solidification reduces the change in volume of the metal from liquid to solid and makes production of more complex castings possible.

The various types of unalloyed cast irons cannot be designated by chemical composition because of similarities between the types. Unalloyed cast irons are designated by their mechanical properties. High-alloy cast irons are designated by their chemical composition and mechanical properties. These have a wide range in chemical composition and also contain major quantities of other elements.

The presence of certain minor elements is also vital for the successful production of each type of cast iron. For example, nucleating agents, called inoculants, are used in the production of gray cast iron to control the graphite type and size. Trace amounts of bismuth and tellurium in the production of malleable cast iron, and the presence of a few hundredth of a percent of magnesium causes the formation of spheroidal graphite cast iron.

In addition, the composition of a cast iron must be adjusted to suit particular castings. Small castings and large castings of the same grade of cast iron cannot be made from the same composition of alloy. For this reason, most cast iron castings are purchased on the basis of mechanical properties rather than composition. The common exception is for castings that require special properties such as corrosion resistance or elevated temperature strength.

The various types of cast iron can be classified according to their microstructure. This classification is based on the form and shape in which the major portion of carbon occurs in the cast irons. This system provides for five basic types of gray cast iron, spheroidal graphite cast iron, malleable cast iron, compacted graphite cast iron, and white cast iron. Each of these types may be moderately alloyed or heat treated without changing its basic classification. The high-alloyed cast irons, generally containing over 3% of added alloying element, can also be individually classified as gray or spheroidal graphite cast iron or white cast iron, but the high-alloyed cast irons are classified as a separate group.

Gray Cast Iron. When the composition of a molten cast iron and its cooling rate are appropriate, the carbon in the cast iron separates during solidification and forms separate graphite flakes that are interconnected within each eutectic cell (EN 1561). The graphite grows edgewise into the liquid and forms the characteristic flake shape. When gray cast iron is broken, most of the fracture occurs along the graphite, thereby accounting for the characteristic gray color of the fractured surface.

Because the large majority of iron castings produced are of gray cast iron, the generic term is often improperly used to mean gray cast iron specifically.

The properties of gray cast iron are influenced by the size, amount, and distribution of the graphite flakes, and by the relative hardness of the matrix metal around the graphite. These factors are controlled mainly by the carbon and silicon contents of the metal and the cooling rate of the casting. Slower cooling and higher carbon and silicon contents tend to produce more and larger graphite flakes, a softer matrix structure, and lower strength. The flake graphite provides gray cast iron with unique properties such as excellent machinability at hardness levels that produce superior wear-resisting characteristics, the ability to resist galling, and excellent vibration damping.

The amount of graphite present, as well as its size and distribution, are important to the properties of the cast iron. Wherever possible, it is preferable to specify the desired properties rather than the factors that influence them.

Microscopically, all gray cast irons contain flake graphite dispersed in a iron-silicon matrix. How much graphite is present, the length of the flakes, and how they are distributed in the matrix directly influence the properties of the cast iron.

The basic strength and hardness of the cast iron is provided by the metallic matrix in which the graphite occurs. The properties of the metallic matrix can range from those of a soft, low-carbon steel to those of hardened, high-carbon steel. The matrix can be entirely ferritic for maximum machinability, but the cast iron will have reduced wear resistance and strength. An entirely pearlitic matrix is characteristic of high-strength gray cast iron, and many castings are produced with a matrix microstructure of both ferrite and pearlite to obtain intermediate hardness and strength. Alloying element additions and/or heat treatment can be used to produce gray cast iron with very fine pearlite or with an acicular matrix structure.

Graphite has little strength or hardness, so it decreases these properties of the metallic matrix. However, graphite provides several valuable characteristics to gray cast iron: the ability to produce sound castings economically in complex shapes, good machinability, even at wear-resisting hardness levels and without burring, dimensional stability under different heating, high vibration damping, and borderline lubrication retention.

The properties of gray cast iron primarily depend on its chemical composition. The lower strength grades of gray cast iron can be produced consistently by simply selecting the proper melting stock. Grey cast iron castings in the higher strength grades require close control of their processing and chemical composition.

The majority of the carbon in gray cast iron is present as graphite. Increased amounts of graphite result from an increased total carbon content in the gray cast iron. This decreases the strength and hardness of the gray cast iron, but increases other desirable characteristics.

An appreciable silicon content is necessary in gray cast iron because this element causes the precipitation of the graphite in the material. The silicon also contributes to the distinctive properties of the gray cast iron. It maintains a moderate hardness level, even in the fully annealed condition, and thus assures excellent machinability. Also, silicon imparts corrosion resistance at elevated temperature and oxidation resistance in gray cast iron.

Gray cast iron can be alloyed to increase its strength and hardness as-cast or its response to hardening by heat treatment.

A very important influence on gray cast iron properties is the effective section thickness in which it is cast. The thicker the wall and the more compact the casting, the lower the temperature at which liquid metal will solidify and cool in the mold. As with all metals, slower solidification causes a larger grain size to form during solidification. In gray cast iron, slower solidification produces a larger graphite flake size.

Gray cast iron is commonly classified by its minimum tensile strength or by hardness (Table 7.1). The mechanical properties of gray cast iron are determined by the combined effect of its chemical composition, processing technique in the foundry, and the solidification and cooling rates. Thus, the mechanical properties of the gray cast iron in a casting will depend on its shape, size and wall thickness as well as on the gray cast iron that is used to pour it. Five grades of gray cast iron are classified by their tensile strength in EN 1561. The grades of gray cast iron also can be specified by Brinell hardness only. The chemical composition and heat treatment, unless specified by the purchaser, shall be left in the direction of the manufacturer, who shall ensure that the casting and heat treatment process is carried out with the same process parameters.

Spheroidal Graphite Cast Iron. Spheroidal graphite cast iron or ductile iron (EN 1563) is characterized by the fact that all of its graphite occurs in microscopic spherolites. Although this graphite constitutes about 10% by volume of this material, its compact spheroidal shape minimizes the effect on mechanical properties.

The difference between the various grades of spheroidal graphite cast irons is in the microstructure of the material around the graphite, which is called the matrix. This microstructure varies with the chemical composition and the cooling rate of the casting. It can be slowly cooled in the sand mold for a minimum hardness as-cast condition or, if the casting has sufficiently uniform sections, it can be shaken out of the mold while still at a temperature above the critical and normalized.

The matrix microstructure and hardness can also be changed by heat treatment. The high ductility grades are usually annealed so that the matrix structure's ferrite is entirely free of carbon. The intermediate grades are often used in the as-cast condition without heat treatment and have a matrix structure of ferrite and pearlite. The ferrite occurs as rings around the graphite spheroids. Because of this, it is called bull-eye ferrite. The high-strength grades are usually given a normalizing heat treatment to make the matrix all pearlite, or they are quenched and tempered to form a matrix of tempered martensite. However, spheroidal graphite cast iron can be moderately alloyed to have an entirely pearlitic matrix as-cast condition.

The chemical composition of spheroidal graphite cast iron and the cooling rate of the casting directly affect its tensile properties by influencing the type of matrix structure that is formed. All of the regular grades of the spheroidal graphite cast iron can be made from the same cast iron provided that the chemical composition is appropriate so that the desired matrix microstructure can be obtained by controlling the cooling rate of the casting after it is poured or by subsequent heat treatment. For most casting requirements, the chemical composition of the spheroidal graphite cast iron is primarily a matter of facilitating production.

Table 7.1 Mechanical properties of gray cast iron

Tensile strength (N/mm^2)	100–350
Brinell hardness	155–265

Table 7.2 Mechanical properties of spheroidal graphite cast iron

Tensile strength (N/mm^2)	350–900
Yield strength (N/mm^2)	220–600
Elongation (%)	6–22
Brinell hardness	130–330

The common grades of spheroidal graphite cast iron differ primarily in the matrix structure that obtains the spheroidal graphite. These differences are the result of differences in the chemical composition, in the cooling rate of the casting, or the result of heat treatment.

13 grades of spheroidal graphite cast iron are classified by their tensile properties or hardness in EN 1563 (Table 7.2). The common grades of spheroidal graphite cast iron also can be specified by only Brinell hardness. The method of producing spheroidal graphite, the chemical composition and heat treatment unless will be specified by the purchaser.

Ausferrite Spheroidal Graphite Cast Iron. This group of spheroidal graphite cast iron (ISO/WD 17804, ASTM 897-90) is well known as ADI (austempered cast iron), and recently as ausferrite spheroidal graphite cast iron. The development of ausferrite spheroidal graphite cast iron has given the design engineer with a new group of cast ferrous materials that offer the exceptional combination of mechanical properties equivalent to cast and forged steels and production costs similar to those of conventional spheroidal graphite cast iron.

Ausferrite spheroidal graphite cast iron provides a wide range of properties, all produced by varying the heat treatment (austempering) of the same castings. Austempering is a special heat treatment process, which consists of three steps:

- Austenitize in the temperature range of 840–950 °C for a time sufficient to produce a fully austenitic matrix that is saturated with carbon.
- Rapidly cool the entire part to an austempering temperature in the range of 230–400 °C without forming pearlite or allowing the formation of ausferrite to begin.
- Isothermally treat at the austempering temperature to produce ausferrite with an austenite carbon content in the range of 1.8–2.2%.

After heat treatment (austempering) the matrix consists of acicular ferrite and residual austenite without carbides.

Six grades of ausferrite spheroidal graphite cast iron are classified by their tensile properties in ISO/WD 17804 and two abrasion resistant grades are classified by Vickers hardness.

Compacted Graphite Cast Iron. Compacted graphite cast iron (vermicular graphite cast iron, VDG-sheet W50) is a recent addition to the family of commercially produced cast irons (Table 7.4). Its characteristics are between of the gray cast iron and spheroidal graphite cast iron. The graphite in compacted graphite cast iron is in the form of interconnected flakes. The short span and blunted edges of graphite in this material provide improved strength, some ductility and a better machined finish than gray cast iron. The interconnected compacted graphite cast iron provides slightly higher thermal conductivity, more damping capacity, and better machinability than those obtained with spheroidal graphite cast iron.

Compacted graphite cast iron provides similar tensile and yield strengths to ferritic spheroidal graphite cast iron and malleable cast iron, although the ductility is less.

Table 7.3 Mechanical properties of ausferrite spheroidal graphite cast iron

Tensile strength (N/mm^2)	800–1400
Yield strength (N/mm^2)	500–1100
Elongation (%)	1–10
Brinell hardness	250–480
Abrasion resistant spheroidal graphite ausferritic cast irons	
Tensile strength (N/mm^2)	1400–1600
Yield strength (N/mm^2)	1100–1300
Elongation (%)	0–1
Vickers hardness	400–500

Table 7.4 Mechanical properties of compacted graphite cast iron

Tensile strength (N/mm^2)	300–500
Yield strength (N/mm^2)	220–380
Elongation (%)	0.5–1.5
Brinell hardness	140–260

Malleable Cast Iron. The starting point is a cast iron in which the carbon and silicon contents are arranged so that the casting is graphite-free after solidification, the entire carbon content being bonded to the iron carbide (cementite). If the casting is then heat-treated (tempered), the cementite decomposes without residue.

Two kinds of malleable cast iron are distinguished:

- White malleable cast iron, which is decarbonized during heat treatment; and
- black malleable cast iron, which is not decarbonized during heat treatment

White malleable cast iron (EN 1562) is produced by heating for 50–80 h at about 1050 °C in a decarburizing atmosphere (CO, CO_2, H_2, H_2O). In this process carbon

Table 7.5 Mechanical properties of malleable cast iron

White malleable cast iron	
Tensile strength (N/mm^2)	350–500
Yield strength (N/mm^2)	170–350
Elongation (%)	3–16
Brinell hardness	200–250
Black malleable cast iron	
Tensile strength (N/mm^2)	300–800
Yield strength (N/mm^2)	200–600
Elongation (%)	1–10
Brinell hardness	150–320

Table 7.6 Mechanical properties of austenitic cast iron

With flake graphite	
Tensile strength (N/mm^2)	140–220
Elongation (%)	2
Brinell hardness	120–150
With spheroidal graphite	
Tensile strength (N/mm^2)	370–500
Yield strength (N/mm^2)	210–290
Elongation (%)	1–45
Brinell hardness	120–255

is removed from the casting, so that after cooling a purely ferritic microstructure is in the casting. White malleable cast iron with small cross sections is welds well.

Black malleable cast iron is produced by heating in a neutral atmosphere, first for about 30 h at 950 °C. In this process the cementite of the ledeburite decomposes into austenite and graphite (temper carbon), which is precipitated in fluky clusters. In a second step of the heat treatment the austenite is converted during slow cooling from 800 to 700 °C into ferrite and temper carbon or transformed during quick cooling into pearlite and temper carbon.

In EN 1562 five grades of white malleable cast iron and nine grades of black malleable cast iron are classified by tensile strength.

Austenitic Cast Iron. High-alloy cast iron is used to produce components that require resistance to corrosives in the operating environment such as seawater, sour well oils, commercial organic and inorganic acids, and alkalis. The ability to easily cast it into complex shapes and the ease of machining some types of this material, make high-alloy cast iron an attractive material for the production of components for chemical processing plants, petroleum refining, food handling, and marine service. Two types dominate high alloy corrosion resistant cast iron: nickel-alloyed cast iron (austenitic cast iron, EN 13835) and high-Si cast iron.

Nickel-alloyed cast iron owes its excellent corrosion resistance to the presence of nickel in concentrations of 12.0–36.0%, a chromium content of 1.0–5.5% and, in one type, s copper content of 5.5–7.5%. These cast irons have an austenitic matrix.

Ten grades of austenitic cast iron with spheroidal graphite and two grades of austenitic cast iron with flake graphite are classified in EN 13835 by chemical composition and mechanical properties, like austenitic steels.

The mechanical properties of austenitic cast iron with spheroidal graphite and with flake graphite are shown in Table 7.6.

Abrasion Resisting Alloyed Cast Iron. High-alloy white cast iron (EN 12513) is specially qualified for abrasion-resistant applications. The predominant carbides in its microstructure provide the high hardness necessary for crushing and grinding other materials without degradation. The supporting matrix structure may be adjusted by alloy content and/or heat treatment to develop the most cost-effective balance between resistance to abrasive wear and the toughness required to withstand repeated impact loading. High-alloy white cast iron is easily cast into shapes required for crushing and grinding or the handling of abrasive materials.

Abrasion resistance concerns the conditions under which a metal or alloy is used. The ability of a part to resist a weight loss due to abrasion depends upon its microstructure, the actual mechanical operations of the part, and the kind and size of material being moved, crushed or ground.

Most of the white cast iron designated for abrasion-resistant applications falls within the high-alloy cast iron category, but unalloyed white cast iron is common and provides satisfactory service where the abrading material is not fine or where replacement is not frequent or expensive. All alloyed cast iron contains chromium to prevent the formation of graphite and to ensure the stability of the carbides in the microstructure. Alloy white cast iron also may contain nickel, molybdenum, copper, or a combination of these alloying elements to prevent or minimize the formation of pearlite in the microstructure.

Unalloyed white cast iron castings develop hardnesses in the range 350–550 BHN. Their microstructures consist of primary iron carbides with a microhardness of 900–1200 VHN in a pearlitic matrix with a microhardness of 220–300 VHN. Alloyed martensitic white cast iron, however, develops Brinell hardnesses in the 500–700 range. The carbide hardness remains 900–1200 VHN, but martensitic, always associated

with some retained austenite, exhibits a microhardness of 600–700 VHN. For many abrasion-resistant applications; the more costly alloyed white cast iron with martensitic matrix structures provide the most economical service.

EN 12513 covers the composition and hardness of abrasion-resistant white cast iron. Martensitic white cast iron falls into two major groups:

- The low-chromium group with 1–4% chromium and 3–5% nickel
- The high-chromium white cast iron containing 14–28% chromium with 1–3% of molybdenum, often alloyed further with additions of nickel and copper

A third but minor category comprise the straight 25–28% chromium white cast iron.

Cast Steel

Cast steels can be classified into four groups:

- Cast carbon and cast low-alloy steel
- Cast high-alloy steel
- Cast stainless steel
- Cast heat-resisting steel

Cast Carbon and Cast Low-Alloy Steel. This group of cast steels consists of many subgroups: steel castings for general purposes (DIN 1681, EN 10293 steel casting for general engineering uses), steel casting for pressure purposes (partially EN 10213), steel castings with improved weldability and toughness for general purposes (DIN 17182, EN 10293 steel castings for general engineering uses, draft), quenched and tempered steel castings for general purposes (DIN 17205, EN 10293 steel castings for general engineering uses, draft), steel castings for use at room temperature and elevated temperatures (EN 10213-2), and steel castings for use at low temperatures (EN 10293-3).

Carbon steel is considered to be steel in which carbon is the principal alloying element. Other elements that are present and that, in general, must be reported are manganese, silicon, phosphorus, and sulfur. In a sense, all of these elements are *residuals* from the raw materials (coke, iron ore) used in the manufacture of the steel, although the addition of manganese is often made during the steelmaking process to counter the deleterious effect of sulfur.

Low-alloy cast steels are considered to be those steels to which elements (other than carbon) are added deliberately to improve mechanical properties.

For all cast carbon and cast low-alloy steels, the mechanical properties are controlled by the chemical composition, the heat treatment and the microstructure of these cast steels. Among the exceptions are the effect of carbon on increasing hardness, the effect of nickel on increasing toughness, and the effect of combinations of chromium, molybdenym, vanadium, and tungsten on increasing elevated temperature strength. The major reason for using alloying elements in low-alloy cast steels is to make the role of heat treatment on increasing strength effective over a wide range of material thickness by quenching and tempering. This effectiveness is termed *hardenability*.

30 grades of steel castings for general engineering uses (5 grades of carbon cast steels, 20 grades of low alloy cast steels, and 5 grades of high alloy cast steels) are classified by their chemical composition, heat treatment processes (austenitizing, air cooling/austenitizing, quenching, tempering), and mechanical properties in EN 10293 (Table 7.7).

EN 10213 consists of steel castings for pressure purposes, in specially cast steel grades for use at room temperature and elevated temperatures (carbon cast steels, low alloy cast steels, high alloy cast steels), cast steel grades for use at low temperatures (low alloy cast steels, high alloy cast steel), and cast austenitic and austenitic-ferritic steel grades (high alloy cast steel grades).

High-Alloy Cast Steel. There are two main groups of high-alloy cast steels: cast stainless steels and cast heat-resisting steels.

Cast Stainless Steel. Cast stainless steels (EN 10213, EN 10283, SEW 410) are distinguished by special resistance to chemically corrosive substances; in general, they have a chromium content of at least 12 wt %. The cast stainless steels in EN 10213, EN 10283, and SEW 410 are subdivided into martensitic, ferritic-carbidic, ferritic-austenitic, austenitic, and full austenitic steels. Cast stainless steels are suitable for welding. Their resistance to intercrystalline corrosion in mill finish is an important property of cast stainless steels.

A special kind of cast stainless steels are the duplex-steels (dual phase steels) with about 50% austenite and 50% soft martensite, in which the two phases fulfil different functions: the austenite guarantees corrosion

Table 7.7 Mechanical properties of steel castings for general engineering uses

Tensile strength (N/mm^2)	380–1250
Yield strength (N/mm^2)	200–1000
Elongation (%)	7–25

protection, e.g. seawater resistance in this case, the soft martensite guarantees component strength.

In EN 10213, EN 10283, and SEW 410 44 grades of cast stainless steels are classified by chemical composition and mechanical properties. The main alloying elements are chromium, nickel, and molybdenum.

Heat-Resisting Cast Steel. The chief requirement for heat-resisting cast steels (EN 10295, SEW 471, SEW 595) is not especially good high-temperature strength but sufficient resistance to hot gas corrosion in the temperature range above 550 °C. The highest temperature at which a heat-resisting steel can be used depends on operational conditions. Recommended temperatures for air and hydrogen atmospheres are up to 1150 °C depending on the chemical composition.

The scaling limit temperatures for the heat-resisting steels is defined as the temperature at which the material loss in clean air is 0.5 mg cm^{-2} h^{-1}.

The scale resistance of heat-resisting cast steels is based on the formation of dense, adhesive surface layers of oxides of the alloying elements chromium, silicon, and aluminum. The protective effect starts when the chromium content is 3 to 5%, but chromium contents up to 30% can be alloyed. The protective effect of these layers is limited by the corrosive low-melting-point eutectics and by carburizing. To increase the heat resistance the alloying element nickel is added in addition to chromium (Cr + Ni = 25–35%).

In EN 10295, SEW 471, and SEW 595 25 grades of heat-resistant cast steels are classified by the chemical composition and the mechanical properties. The main alloying elements are chromium, silicon, and nickel.

The creep behavior with the creep rupture strength and the creep limit in the temperature range of 600 up to 1100 °C is the most important.

Cast Nonferrous Alloys

The cast nonferrous alloys are classified into four main groups:

- Cast aluminum alloys
- Cast magnesium alloys
- Cast copper alloys
- Cast zinc alloys

There are other groups: for example, cast titanium alloys, cast tin alloys, cast lead alloys, cast nickel alloys, cast cobalt alloys, etc.

Cast Aluminum Alloys. The specification of a cast aluminum alloy (EN 1706) for a cast component is based upon the mechanical properties it can achieve. These properties are obtained from one particular combination of cast alloy, melt treatment (grain refining, modification) foundry practice, and thermal treatment. In all cast aluminum alloys the percentage of alloying elements and impurities must be carefully controlled.

The main alloying elements of the cast aluminum alloys are copper, silicon, magnesium, and zinc. Grain refiners, which are usually materials that liberate titanium, boron, or carbon, are generally added in the form of master-alloy to the melt. In casting alloy this is a well-proven method to influence the nucleation conditions in a melt, so that it solidifies with as fine-grained and dense a structure as possible. Hypereutectic aluminum-silicon alloys can be grain-refined with additions that release phosphorus, which promotes the nucleation of primary silicon. Modifying aluminum-silicon alloys of eutectic and hypereutectic composition means treating the melt to binder primary silicon from precipitating to form coarse, irregularly shaped particles. The melt can be modified by adding capsules of metallic sodium or compounds that release sodium. Alternatively, the addition of strontium has proved successful in castings. In contrast to sodium, which burns off and is lost fairly quickly, strontium lasts longer.

Industrial casting processes consist of traditional sand casting, low-pressure sand casting, investment casting, lost-foam casting, permanent mold casting, high pressure die casting, low-pressure permanent mold casting, back-pressure die casting, vacuum die casting, squeeze casting, and thixocasting.

Sand and permanent mold castings may be heat treated to improve mechanical and physical properties. The following thermal treatments are industrially used:

- Stress relief or annealing
- Solution heat treatment and quenching, artificial aging

Table 7.8 Mechanical properties of cast stainless steels

Tensile strength (N/mm^2)	430–1100
Yield strength (N/mm^2)	175–1000
Elongation (%)	5–30

Table 7.9 Mechanical properties of heat-resistant cast steel at room temperature

Tensile strength (N/mm^2)	400–440
Yield strength (N/mm^2)	220–230
Elongation (%)	5–15

Table 7.10 Mechanical properties of aluminum cast alloys (sand molding – 1, permanent mold casting – 2, high pressure die casting – 3, investment casting – 4)

Casting technology	1	2	3	4
Tensile strength (N/mm^2)	140–300	150–330	200–240	150–300
Yield strength (N/mm^2)	70–210	70–280	120–140	80–240
Elongation (%)	1–5	1–8	1–2	1–5
Brinell hardness	40–100	45–100	55–80	50–90

- Solution heat treatment, quenching and natural aging and
- Solution heat treatment, quenching, and artificial overaging (for the groups 1, 2 and 4 in Table 7.10)

In EN 1706 37 grades of cast aluminum alloys are classified by their chemical composition and mechanical properties (Table 7.10). The mechanical properties depend on the chemical composition of the cast aluminum alloys, the casting technology, and the heat treatment process.

Cast Magnesium Alloys. Magnesium combines a density two-thirds that of aluminum and only slightly higher than that of fiber-reinforced plastics with excellent mechanical and physical properties as well as processability and recyclability.

Cast magnesium alloys (EN 1753) can be divided into two groups: the sand-casting alloys that have a fine grain structure due to a melt treatment with small additions of zirconium, and the die casting alloys, in which aluminum is the principal alloying element. The alloys can also be classified as general purpose, high ductility, and high temperature alloys. Most of the alloys are produced as high-purity versions to reduce potential corrosion problems associated with higher levels of iron, nickel, and copper.

Aluminum improves the mechanical strength, corrosion properties, and castability of the castings. Ductility and fracture toughness are gradually reduced with increasing aluminum content.

Manganese is added to control the iron content of the alloys. The level of manganese additions varies from one alloy to the next, depending on the mutual solubilities of iron and manganese in the presence of other alloying elements. A basic requirement of high-purity alloys is that the iron content of diecast parts is limited to a maximum of 0.005 wt %. Other impurities like nickel and copper also must be strictly controlled. Other alloying elements are zinc, manganese, silicon, copper, zirconium, and rare earth elements.

Following are some of the advantages magnesium alloys offer casting designers:

- *Light weight* – The lightest of all structural alloys, magnesium alloys preserves the light weight of a design without sacrificing strength and rigidity.
- *High stiffness to weight ratio* – This characteristic is important where resistance to deflection is desired in a light-weight component.
- *Damping capacity* – Magnesium is unique among metals because of its ability to absorb energy inelastically. This property yields the vibration absorption capacity to ensure quieter operation of equipment.
- *Dimensional stability* – Annealing, artificial aging or stress-relieving treatments normally are not necessary to achieve stable final dimensions.
- *Impact and dent resistance* – The elastic energy absorption characteristics of magnesium alloys result in a good impact and dent resistance and energy management.
- *Anti-galling* – Magnesium alloys possess a low galling tendency and can be used as a bearing surface in conjunction with shaft hardness above 400 HB.
- *High conductivity* – Magnesium alloys have a high thermal conductivity and a good electrical conductivity.
- *Wall thickness* – Magnesium alloy die castings are commonly produced with a wall thickness from 0.15 to 0.4 cm.

Magnesium alloys can be cast by a variety of methods, including high-pressure die casting, low pressure permanent mold casting, sand casting, plaster/investment casting, and thixocasting and squeeze casting.

Different alloys may be specified for the different processes. In cases where the same alloy is used with different casting processes, it is important to note that the properties of the finished castings will depend on the casting method. The most prevalent casting method

Table 7.11 Mechanical properties of cast magnesium alloys

	Casting method Sand casting	Permanent mould casting	High-presure diecasting
Tensile strength (N/mm^2)	140–250	160–250	150–260
Yield strength (N/mm^2)	90–175	90–175	80–160
Elongation (%)	2–8	2–8	1–18
Brinell hardness	50–90	50–90	50–85

Table 7.12 Mechanical properties of cast copper alloys

Tensile strength (N/mm^2)	150–750
Yield strength (N/mm^2)	40–480
Elongation (%)	5–25
Brinell hardness	40–190

Table 7.13 Mechanical properties of cast zinc alloys

Tensile strength (N/mm^2)	220–425
Yield strength (N/mm^2)	200–370
Elongation (%)	2.5–10
Brinell hardness	83–120

for magnesium alloys is die casting. In this process, thin-walled parts are produced at high production rates with reduced tool wear compared to aluminum alloys, due to the lower heat content per volume of molten metal. Both hot chamber and cold camber machines are currently used for magnesium alloys. Thixocasting is another casting method that has shown progress with magnesium alloys.

There are seven cast magnesium alloys in EN 1753.

Cast Copper Alloys. Cast copper alloys (EN 1982) are known for their versatility. They are used in a wide range of applications because they are easily cast, have a long history of successful use, are readily available from a multitude of sources, can achieve a range of physical and mechanical properties, and are easily machined, brazed, soldered, polished, or plated.

The following lists the physical and mechanical properties common to cast copper alloys:

- Good corrosion resistance, which contributes to the durability and long-term cost-effectiveness.
- Favorable mechanical properties ranging from pure copper, which is soft and ductile, to manganese-bronze, which rivals the mechanical properties of quenched and tempered steel. In addition, all cast copper alloys retain their mechanical properties, including impact toughness at low temperatures.
- High thermal and electrical conductivity, which is greater than any metal except silver. Although the conductivity of copper drops when alloyed, cast copper alloys with low conductivity still conduct both heat and electricity better than other corrosion-resistant materials.
- Bio-fouling resistance, as copper inhibits marine organism growth. Although this property (unique to copper) decreases upon alloying, it is retained at a useful level in many alloys, such as copper-nickel.
- Low friction and wear rates, such as with the high-leaded tin-bronzes, which are cast into sleeve bearings and exhibit lower wear rates than steel; good castability, as all cast copper alloys can be sand cast and many can be centrifugally, continuously, and permanent mold cast, as well as diecast.
- Good machinability, as the leaded copper alloys are free-cutting at high machining speeds, and many unleaded alloys such as nickel-aluminum bronze are readily machinable at recommended feeds and speeds with proper tooling.
- Ease of post-casting processing, as good surface finish and high tolerance control is readily achieved. In addition, many cast copper alloys are polished to high luster, and platting, soldering, and welding also are routinely performed.
- Large alloy choice, since several alloys may be suitable candidates for any given application depending upon design loads and corrosivity of the environment.
- Comparable costs to other metals due to their high yield, low machining costs, and little requirement for surface coatings such as paint.

In EN 1982 the cast copper alloys are divided into cast copper, cast copper-chromium, cast copper-zinc, cast copper-tin, cast copper-tin-lead, cast copper-aluminum, cast copper-manganese-aluminum, and cast copper-nickel alloys.

35 grades of cast copper alloys are classified by their chemical composition and mechanical properties in EN 1982.

Cast Zinc Alloys. Cast zinc alloys (EN 1774 and EN 12844) are assigned to three alloy groups. The first group of alloys have 4% aluminum as the primary alloying element with 0.099% or less magnesium to control intergranular corrosion. Another alloying element is copper. The alloys with the highest copper content have the highest hardness. The mechanical properties can be improved with 0.005 to 0.2% nickel as alloying element. The second group has higher aluminum contents (8 to 27%) These alloys have superior hardness, wear and creep resistance that increase with the aluminum content. The third group is a cast zinc alloy that has copper as the primary alloying element. Castings of the cast zinc alloys are manufactured by the high pressure die casting.

In EN 12844 8 grades of cast zinc alloys are classified by their chemical composition and mechanical properties.

7.1.4 Primary Shaping

According to DIN 8580 [7.1], primary shaping is the manufacturing of a solid body from a shapeless material by creating cohesion. Thus primary shaping serves to give a component made from a material in shapeless condition an initial form. Shapeless materials are gases, liquids, powders, fibers, chips, granules, solutions, melts, and the like. Primary shaping may be divided into two groups with regard to the form of the products and their further processing:

- Products produced by primary shaping, which will be further processed by forming, severing, cutting, and joining. The final product no longer resembles the original product of primary shaping in form and dimensions, i. e. a further material change in shape and dimensions is accomplished by means of other main groups of manufacturing processes.
- Products produced by primary shaping, which essentially have the form and dimensions of finished components (e.g. machine parts) or end-products, i. e. their shape essentially corresponds to the purpose of the product. The attainment of the desired final form and dimensions usually requires only operations that fall into the main process group *cutting* (machining).

Most powders are produced by primary shaping, whereby the powders are atomized out of the melt, and rapid solidification is followed. From powder, sintering parts are produced as a result of powder metallurgical manufacturing.

The production of cast parts from metallic materials in the foundry industry (castings), from metallic materials in powder metallurgy (sintered parts), and from high-polymer materials in the plastics processing industry has major advantages for economic efficiency.

The production of cast parts is the shortest route from the raw material to the finished part. It bypasses the process of forming and all the associated expense. The final form of a finished component with a mass ranging from a few grams to several hundred tonnes is practically achieved in one direct operation.

The production of cast parts by primary shaping from the liquid state allows the greatest freedom of design. This cannot be achieved by any other manufacturing process.

Primary shaping also enables processing of materials that cannot be achieved by means of other manufacturing methods. The direct route from the raw material to the preform or the end-product results in a favorable material and energy balance.

The continual further development of primary shaping processes increasingly permits the production of components and end-products with enhanced practical characteristics, i. e. cast parts with lower wall thicknesses, lower machining allowances, narrower dimensional tolerances, and improved surface quality.

In the following, primary shaping of metallic materials from the liquid state in foundry technology, of metallic materials from the solid state in powder metallurgy, and of high-polymer materials (plastics) from the plasticized state or from solutions is discussed on a common basis with regard to the fundamental technological principles. The discussion is restricted to subjects relevant to mechanical engineering.

For a better appreciation of the relevant principle of action, many detailed technological operations are omitted, which although vital to the specific manufacturing technology, are of minor importance. Furthermore, when discussing the specific primary shaping processes, only products with a simple form are referred to, because the diversity of the possible geometric forms cannot be described here.

Only the most important primary shaping processes are selected, as the large number of technological processes and process variables means that it is impossible to provide anything like a complete description. The

processes are selected first according to their technical importance and second according to the principle of action.

Materials technology problems will only be mentioned briefly, although they are vital in order to understand the technological processes, their applicability and efficiency, and the changes in material properties brought about by the technological processes.

Process Principle in Primary Shaping

In the processes of primary shaping, the technological manufacturing process essentially comprises the following steps:

- Supply or production of the raw material as an amorphous substance
- Preparation of a material state ready for primary shaping
- Filling of a primary shaping tool with the material in a state ready for primary shaping
- Solidification of the material in the primary shaping tool
- Removal of the product of primary shaping from the primary shaping tool

These individual steps are discussed in the following section.

Material State Ready for Primary Shaping

In primary shaping of metallic materials from the liquid state, the raw materials (pig-iron, scrap, ferroalloys and the like) are melted in a metallurgical melting furnace by means of thermal energy. The melting furnaces are usually physically separated from the primary shaping tool. The molten metal is carried by means of transfer vessels (ladles) to the primary shaping tools, termed molds in the foundry industry, and cast there.

In primary shaping of high-polymer materials from the plasticized state, bulk raw materials (granules, powder) are fed after proportioning into a preparation device, which is usually integral with the primary shaping tool. There, thorough mixing, homogenizing and plasticizing of the material to be processed are accomplished under the action of heat and pressure. When solutions are used, these are produced in a mixing unit and then poured into the primary shaping tool. In primary shaping of metallic and also high-polymer materials from the solid state, the bulk raw materials (metal powder, plastic powder, or plastic granules) are poured straight into the primary shaping tool, where they sinter, or first become plastic and then solidify under the action of pressure and thermal energy.

Primary Shaping Tools

The primary shaping tool contains a hollow space which, with the allowance for contraction, usually corresponds to the form of the product (unmachined part) to be manufactured, but may be smaller or larger than the resulting unmachined part. Furthermore, primary shaping tools often contain systems of channels (runners) for feeding the material in the state ready for primary shaping. The allowance for contraction corresponds to the dimensional changes that occur in the material to be processed from the moment of solidification to its cooling to room temperature.

In the production of cast parts, a distinction is made between primary shaping tools for once-only use and those for repeated use. Primary shaping tools for once-only use are only used for primary shaping of metallic materials from the liquid state in foundry technology. They are termed expendable or *dead* molds. Only one product (casting) can be manufactured, as the mold is subsequently destroyed. However, primary shaping tools for repeated use (permanent molds) are also used in foundry technology. A larger quantity of cast parts can be produced. The primary shaping technologies for processing of high-polymer materials and powder metallurgy use only primary shaping tools for repeated use. Primary shaping tools for repeated use are usually made of metallic, and more rarely of nonmetallic, materials. Primary shaping tools for once-only use (dead molds) are made with the aid of patterns.

Filling the Primary Shaping Tools

Filling of the primary shaping tools with the material ready for primary shaping may be accomplished by means of the following principles of action: under the influence of gravity, elevated pressure or centrifugal force and by displacement. The material to be processed can be put into the primary shaping tools in solid, pourable form (e.g. powder), as molten metal in the case of metallic materials, or in plasticized condition, as a solution or as a paste in the case of high-polymer materials.

Change of State Ready for Primary Shaping

Shaping into the Solid State of Aggregation. Liquid metallic materials (molten metals) change by crystallization to the solid state of aggregation on cooling owing to the removal of heat.

Thermoplastics are cooled in the primary shaping tool after forming. As a result of temperature reduction, which is accomplished either by heat removal in cooled tools or in downstream equipment

(cooling baths), the plastic mass passes through the following states: plastic–rubberlike–elastic–solid. In setting by cooling, secondary valency bonds are restored. This process is repeatable; therefore thermoplastics cart be restored to the plastic state by reheating.

Thermosetting plastics or thermosets (cross-linkable plastics) are cured after forming by a hardening process. Primary valency bonds form, and the plasticized mass solidifies directly under the effect of heat and/or pressure. The curing is an irreversible chemical process: thermosets disintegrate on reheating without needing to pass through a plastic state. Fundamental chemical reactions during solidification are polymerization, polycondensation, and polyaddition.

In primary shaping of high-polymer materials, if solutions are used then the transformation to the solid state may be accomplished by the physical process of solvent evaporation.

In primary shaping by sintering, a process of shrinkage of the internal and external surface area of a body formed from powder by pressure takes place. Powder particles that are in contact are joined by the formation or reinforcement of bonds (material bridges) and/or by reducing the pore volume; at least one of the material constituents involved remains solid throughout the process. The bonding of the porous pressed body of powder takes place mainly through diffusion mechanisms.

In connection with the description of the technological aspects of primary shaping, further details of the processes, here not described, go to the special literature.

7.1.5 Shaping of Metals by Casting

Manufacturing of Semifinished Products

This group of primary shaping processes involves the production of initial and intermediate products, which are further processed by, for instance, metal forming (plastic deformation).

The Ingot Casting Process. Here, ingots, slabs, wirebars, etc. are produced in permanent molds made of metal (usually cast iron). These products are converted by metal forming (rolling, forging, pressing, wire drawing, etc.) into a semifinished product (sheet, plate, section, wire) that no longer resembles the original ingot in form and dimensions. In ingot casting a distinction is made between top pouring (downhill casting, Fig. 7.3a), where the mold is filled by directly pouring

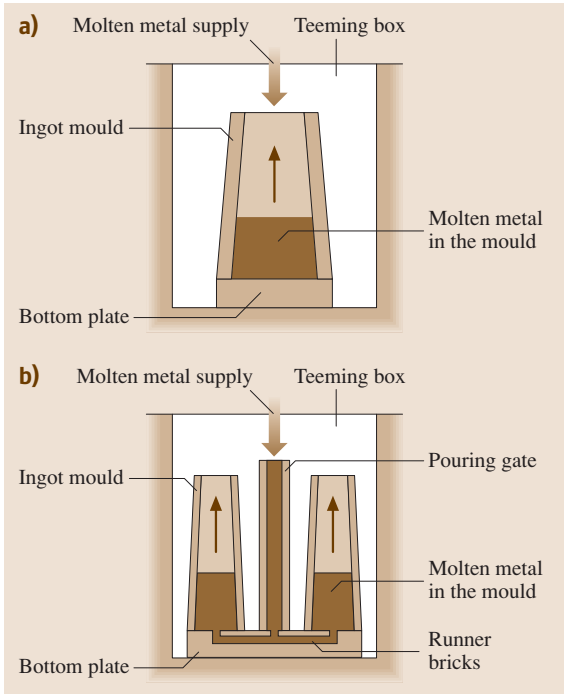


Fig. 7.3a,b Ingot casting methods: (a) top pouring, (b) bottom pouring (after [7.2])

the molten metal in from above, and bottom pouring (uphill casting, Fig. 7.3b), where one mold or several molds simultaneously (group casting) are filled from below by means of a distribution system (pouring gate and runner bricks).

Procedure: The prepared molds are set up into the teeming box as illustrated. They are filled with the liquid metal, which solidifies in them. The molds are stripped from the ingots, which are taken away.

Continuous Casting Processes. In these processes, which are used to produce either intermediate products for metal forming or semifinished products, the primary shaping tool (continuous mold, casting roller, casting belt, casting wheel) is always smaller than the product of primary shaping.

Casting with Stationary Molds. Continuous casting with a stationary mold. In this casting process, a bath of molten metal is fed into a stationary continuous mold, where the solidification begins.

Depending on the design, a distinction is made between batch or continuous vertical (Fig. 7.3a) and horizontal continuous casting systems (Fig. 7.3b). On

leaving the continuous mold the resulting continuous casting (solid or hollow section) is cooled until it solidifies completely. The continuous casting is usually cut into defined lengths at intervals. Like ingots from ingot casting, these are further processed by metal forming.

Traveling Primary Shaping Tools. In this continuous casting process, metal-forming equipment for rolling or drawing is installed directly following the casting plant, thus dispensing with the manufacturing stages of metal forming. In this case, there is usually no cutting of the continuous castings into sections.

Continuous Casting with Moving Molds. The continuous casting with moving molds is realized with strip and wire rod casting plants.

In vertical uphill casting between two casting rollers (Fig. 7.4a) the molten metal is fed from below between two casting rollers. Solidification takes place between these two rollers, and the finished continuous casting (a strip) emerges vertically upward from these rollers.

In horizontal casting (Fig. 7.4b), both the feeding of the molten metal and the discharge of the solidified continuous casting (strip) take place horizontally. In casting between a casting roller or a casting wheel having the profile of the desired strip or rod and an endless casting belt (Fig. 7.5a,b), the molten metal solidifies between the casting roller/wheel and the casting belt and emerges into the open air. In casting in belt molds

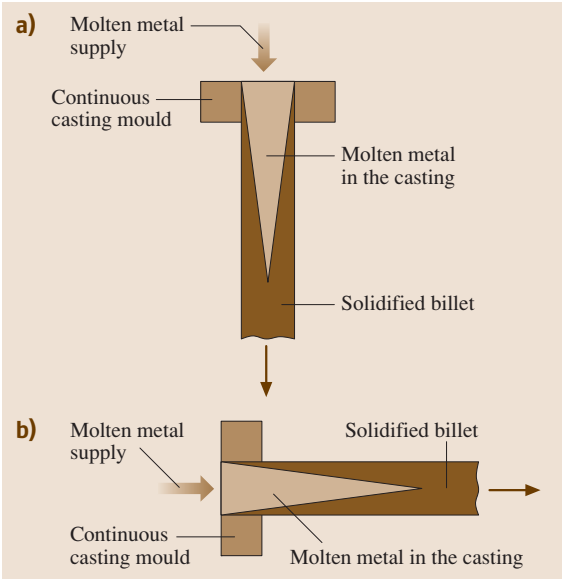


Fig. 7.4a,b Continuous casting: **(a)** vertical, **(b)** horizontal continuous casting (after [7.2])

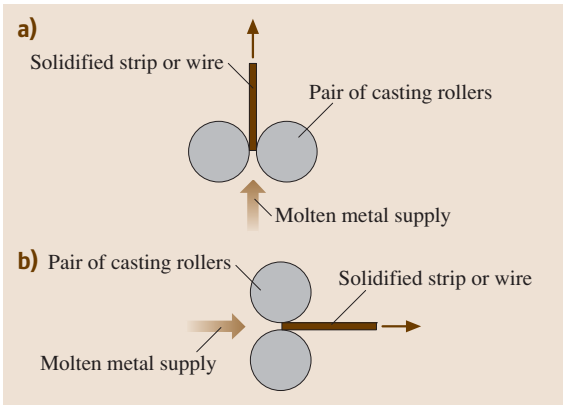


Fig. 7.5a,b Strip casting: **(a)** vertically uphill, **(b)** horizontally (after [7.2])

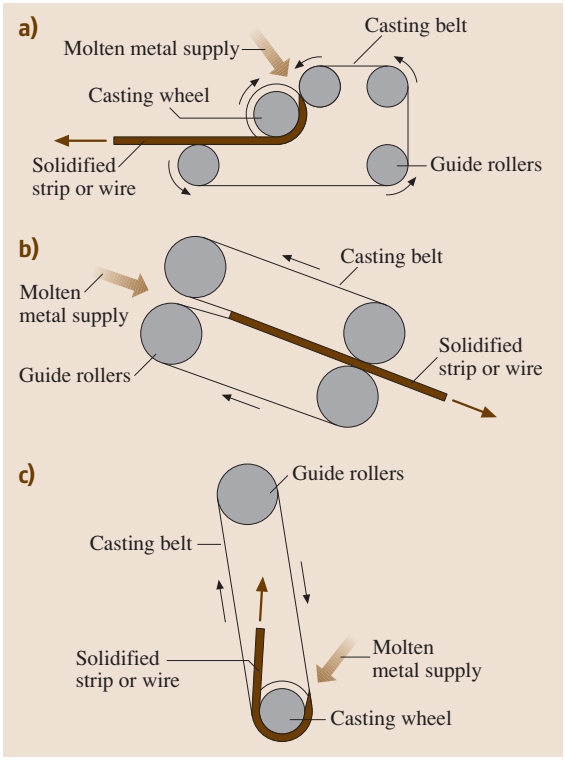


Fig. 7.6a–c Casting machines: **(a)** strip casting machine (rotary process), **(b)** strip casting machine (Hezelett process), **(c)** wire rod casting machine (after [7.2])

(two endless, rotating casting belts), solidification is accomplished with the aid of further rotating equipment to restrict the product laterally between these casting belts (Fig. 7.6a–c); the solidified continuous casting then emerges into the open air as a strip.

Manufacturing of Cast Parts

Manufacturing of cast parts is accomplished with primary shaping processes by means of which a practically finished component, e.g. a machine part or an end-product, is produced without metal forming. The product's shape and dimensions do not undergo any further significant change. However, primary shaping is followed by other manufacturing processes, e.g. cutting (turning, planing, milling, drilling), to obtain a component ready for fitting. The intention is to perfect and further develop the primary shaping techniques in order to, for instance, reduce the amount of machining work to a minimum. A good view of the manufacturing of cast parts – cast alloys, manufacturing processes and applications of cast parts – is given in [7.3, 4].

Use of Lost Primary Shaping Tools (Lost Molds). This technique, which is only used in primary shaping of metallic materials from the liquid state in foundry technology, uses a pattern to produce the expendable primary shaping tool. Depending on the type of pattern used, a distinction is made between processes using a permanent pattern and those using an expendable pattern. A permanent pattern can be used to make many expendable molds, but an expendable pattern can only be used for one expendable mold. Expendable patterns are also made in an appropriate primary shaping tool.

The patterns are similar in shape to the case part to be manufactured, but are larger by the allowance for contraction of the material to be cast. They also incorporate the machining allowances, which will subsequently be eliminated by machining of the casting with the aim of achieving accuracy in dimensions, shape and position, as well as tapers to enable the pattern to be removed from the mold. Most patterns have a pattern joint, i. e. they consist of at least two parts (pattern halves). In addition, for castings with hollow spaces the pattern has care marks for insertion of the cores in the mold.

In the case of permanent patterns for making an expendable mold for casting, these patterns or their sections made from metals, high polymers or wood are used to make the molds by the sand molding, template molding, or shell molding process.

Hand Molding. The mold is expendable (i. e. used only once). The medium used may be natural or synthetic sand with bentonite; other molds used sand with a resin binder. It is worked by hand.

Pattern. Patterns for repeated use, and patterns and core boxes, are made of wood or plastics.

Process Characteristics. Hand molding denotes the production of a sand mold without using a molding machine. The mold consists of the external parts for the external profile and the internal parts for the internal profile. Hollow spaces in the casting are formed by cores placed in the mold.

The principle of molding is illustrated in Fig. 7.7. First of all the bottom half of the two-part pattern is molded. After turning the molding box over, the top half of the pattern and the pouring gate and risers are placed in position and the top mold is made. The top box is lifted off, the pattern halves are removed from the mold, and the core is inserted. The halves of the mold are joined and the casting is made.

Casting Materials. All metals and alloys that are castable with the current technology.

Weight of Castings. The maximum transportable weight and the melting capacity determine the maximum weight.

Number of Castings. Single items, small production runs.

Tolerances. From about 2.5 to 5%.

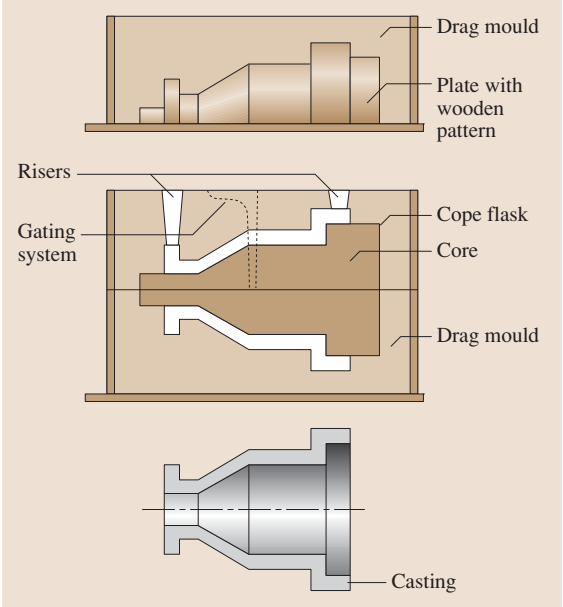


Fig. 7.7 Hand molding (after [7.2])

Machine Molding. The mold is expendable (used only once). Natural sand, artificial sand, sand with synthetic resin binders, CO, sand. Preparation on molding and core molding machines. Used in semiautomatic and fully automatic production lines.

Pattern. Patterns and core boxes are made of metal or plastic.

Process Characteristics. Mechanical molding is characterized by a semiautomatic or fully automatic manufacturing operation for efficient production of ready-to-cast sand molds (Fig. 7.8). The casting process is often incorporated into the production line. The main stages are: molding station, core insertion section, casting section and cooling section. The emptying station releases the cast molds. The molding station may consist of one automatic molding machine for complete molds or of two or more for making separate top and bottom boxes. There are also boxless molding units, where the molds are made using only a frame, which is withdrawn after compacting the sand.

Casting Materials. All metals and alloys that are castable with the current technology.

Weight of Castings. Limited by the size of the molding machines: up to ≈ 500 kg.

Number of Castings. Owing to the mechanical preparation, mechanical molding is suitable for series and mass production of quantities of 1000 and multiples thereof.

Tolerances. From about 1.5 to 3%.

Suction Molding. The mold is expendable (used once only), made from wet artificial casting sand (Fig. 7.9).

Pattern. Wood, plastic, metal.

Process Characteristics. The process is characterized by the formation of a vacuum by withdrawing air from the mold space and the incoming molding sand. This accelerates the sand, which spreads over the wall of the pattern. The sand can be subsequently pressed against the pattern. Advantages of the process are optimum mold, compaction around the pattern, no shadow effect with plane surfaces, decreasing hardness of compacted sand from the inside to the outside, high surface quality, dimensionally stable castings, reduced cleaning. This process should not be confused with vacuum molding.

Casting Materials. Iron, steel, aluminum.

Weights of Castings. From about 0.1 to 120 kg.

Number of Castings. Small, medium, and large production runs.

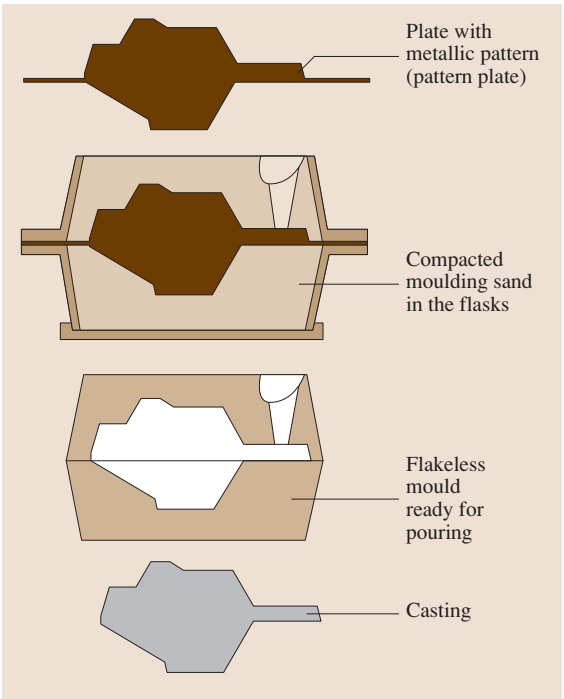


Fig. 7.8 Machine molding (after [7.5])

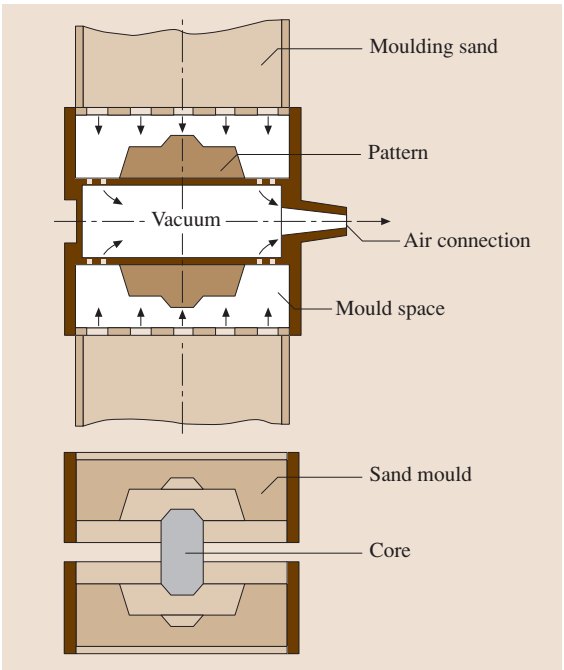


Fig. 7.9 Suction molding (after [7.5])

Tolerances. Conventional to DIN 1683; maximum offset of compacted sand 0.3 mm.

Shell Molding. The mold is expendable (used only once). Resin-coated sands or sand/resin mixtures (Fig. 7.10).

Pattern. Patterns for repeated use, heatable metal patterns, and metal core boxes.

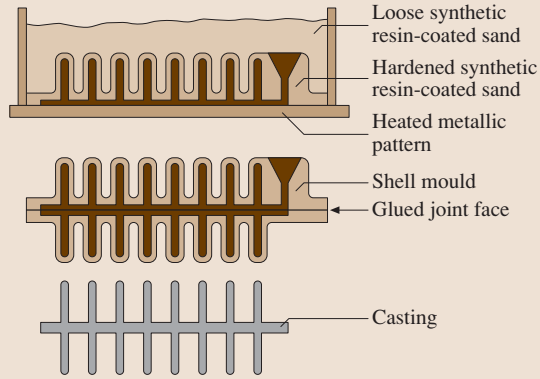


Fig. 7.10 Shell molding (after [7.5])

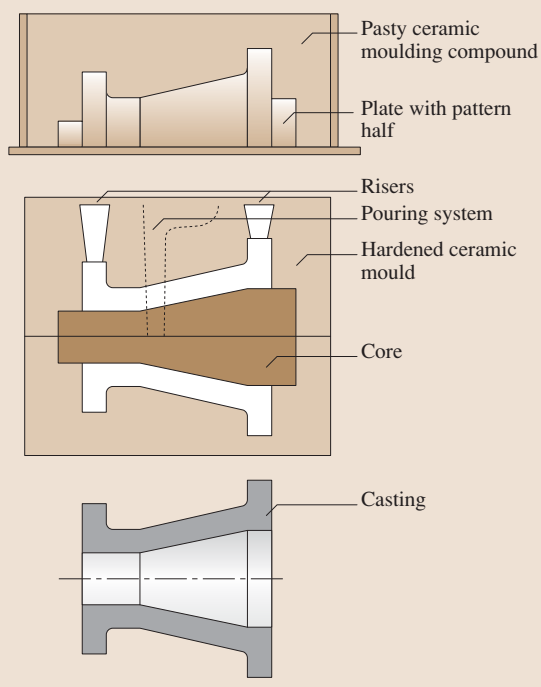


Fig. 7.11 Ceramic molding (after [7.5])

Process Characteristics. These molds are shell molds with walls only a few mm thick. The mold material is poured onto the heated metal pattern. This cures the synthetic resins in the mold material, solidifying the mold. The result is a self-supporting, stable shell mold. Shell molds are often molded in one piece and then divided. After putting in the cores, the two halves of the mold are glued together. The shell molding process is used in various stages of mechanization and automation. This process is used not only for making molds for shell casting, but also for producing hollow shell cores for sand and chill casting. These cores are produced on special core molding machines. Shell casting offers high dimensional accuracy with excellent surface quality.

Casting Materials. All metals and alloys that are castable with the current technology.

Weight of Castings. Up to 150 kg.

Number of Castings. Medium to large production runs.

Tolerances. From about 1 to 2%.

Ceramic Molding. The mold is expendable (used only once). This mold is made of highly refractory ceramic similar in kind to mold materials for investment casting (Fig. 7.11).

Pattern. Reusable, made of metal, plastic, or specially varnished wood.

Process Characteristics. A slip consisting of highly refractory substances is poured around the pattern; these substances then harden by chemical reaction. Often only one layer is poured, which is then back-filled with *normal* molding sand. After removing the pattern, the ceramic is fired or skin-dried (Shaw process). To keep ceramic molding, which is relatively expensive, to a minimum, it is usually only the parts of the mold that are made from special ceramic that will be cast in finished or near-finished shape. Castings from ceramic molds have no casting skin in the conventional sense and are among the precision casting processes that, as the technology develops, are becoming more and more widely used owing to their efficiency.

Casting Materials. All metals and alloys that are castable with the current technology, especially iron-based materials.

Weight of Castings. From about 0.1 to 25 kg, depending on the production equipment.

Number of Castings. Single items, small and medium runs, also larger runs in the case of fluid flow machines.

Tolerances. Up to about 100 mm ± 0.2%, over 100 mm ± 0.3 to 0.8% of nominal dimensions.

Vacuum Molding (V-Process). The mold is expendable (used only once). A plastic foil is vacuum-molded to the contours of the pattern, back-filled with fine-grained, binder-free quartz sand and sealed with a covering foil. Dimensional stability is preserved by means of a vacuum of 0.3 to 0.6 bar.

Pattern. Permanent patterns, not subject to significant wear. Patterns are made of wood or metal. Core boxes according to the core manufacturing method.

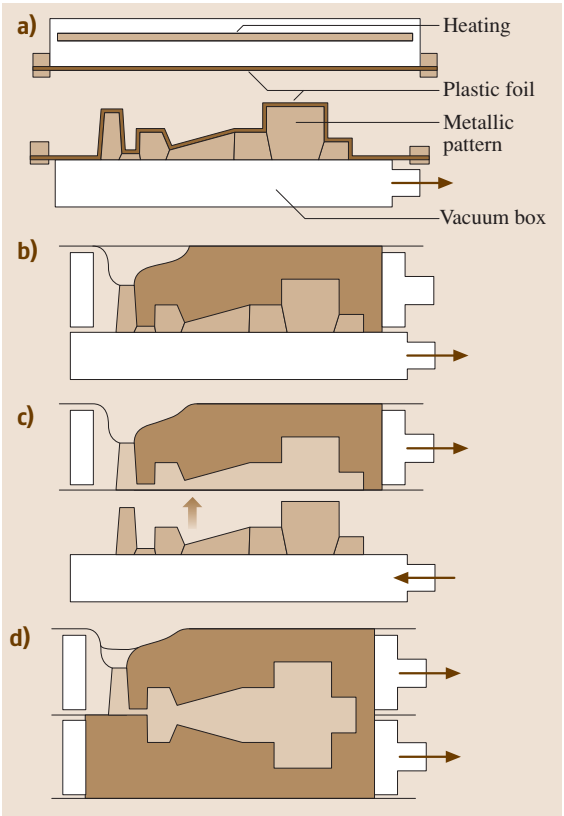


Fig. 7.12a–d Vacuum molding: **(a)** The plastic foil is softened by means of a foil-type heating element and drawn tightly against the pattern by vacuum through holes. **(b)** The mold box is placed on top, filled with binderless sand, precompacted, and the top of the box is covered with plastic foil. **(c)** Vacuum is applied to the mold box, compacting the sand. By switching off the vacuum, the mold box can easily be lifted off the pattern. **(d)** The top and bottom halves of the box are joined. The vacuum is maintained during pouring (after [7.5])

Process Characteristics (Fig. 7.12). The process is characterized by the use of a vacuum for both deep drawing the pattern sheet over a pattern with nozzle holes and maintaining the stability of the mold. A molding box equipped with suction systems is connected by a pipe to the vacuum grid. The fine, binder-free sand with which the mold box is filled is compacted by vibration. After applying a cover sheet, the air is evacuated from the sand and the mold thus becomes rigid. The mold is constantly connected to the vacuum grid before, during, and after casting. To empty the mold, the vacuum is switched off and the sand and cast parts drop out of the mold box without additional force. The advantages of the process are: high, reproducible dimensional accuracy with outstanding surface quality; the mold seam at the joins and core marks is very small; tapers can be entirely dispensed with in certain areas of the casting.

Casting Materials. All metals and alloys that are castable with the current technology.

Weight of Castings. Restricted by the equipment available, not by the process.

Tolerances. 0.3 to 0.6%.

Casting Under Vacuum. The mold is expendable (used once only), shell molds (for investment casting) and precision casting molds made of special mold materials.

Pattern. Wax for investment casting, also of metal, plastic or the like, depending on the type of mold.

Process Characteristics. Titanium and zirconium are among the reactive metals that have high affinities with oxygen, nitrogen, and hydrogen in the molten state. This is even the case when present as alloying constituents in appropriate percentages in, e.g., molten nickel. All these alloys must therefore be produced and cast under defined conditions, normally under high vacuum.

The new mold ceramics, e.g. those made of yttrium and zirconium oxides, resist attack by reactive metals and melts. However, these special ceramics are not (yet) required for nickel-base alloys that are only alloyed with titanium, aluminum, etc.

To optimize quality and structure, the castings are usually isostatically pressed at high temperature by means of the HIP process.

Casting Materials. Alloys based on (in order of importance) nickel, titanium, cobalt, iron, and zirconium.

Weight of Castings. Approximately 0.01 to 100 kg and more, depending on the manufacturing equipment.

Number of Castings. Small series to fairly large production runs.

Tolerances. From 0.3 to ±0.8% of nominal size, depending on the molding process.

Investment Casting (Lost Wax Process). The mold is expendable (used once only), made of highly refractory ceramic, single or group pattern with runners, combined to form casting units (*clusters* or *trees*) (Fig. 7.13)

Pattern. Made by injection molding from special waxes or the like, thermoplastics or mixtures thereof.

Process Characteristics. The distinguishing features are the expendable patterns, the one-piece molds and the casting in hot molds (900 °C for steel). A casting skin in the conventional sense does not form. The patterns are inject ion-molded in single or multiple tools made of aluminum, steel or soft metal, for which an original pattern is required. The most suitable injection molding tool in each particular case is chosen according to the planned total quantity, the form of the casting, and the nature of the pattern material. The formation of certain undercut contours may require the use of water-soluble or ceramic cores, for which a supplementary tool is used. The patterns are assembled into clusters by means of casting systems, usually again employing injection molding. The method of this assembly is crucial for the quality of the castings and for efficiency. Viscous ceramic coatings which cure by chemical reaction are then applied to these clusters. For aluminum, special plasters are also used. After melting out (lost wax process) or dissolving away the pattern material, the resulting one-piece molds are fired. Casting takes place in molds that are usually still hot from firing, so that narrow cross-sections and fine profiles *turn out* cleanly. Precision casting, with its tight tolerances and high surface quality, is the casting technology that offers the greatest freedom of design coupled with high quality.

Casting Materials (in Order of Importance). Steels and alloys based on iron, aluminum, nickel, cobalt, titanium, copper, magnesium, or zirconium, including aerospace, materials, produced at atmospheric pressure or under vacuum.

Weight of Castings. 0.001 to 50 kg, also up to 150 kg and over depending on manufacturing equipment.

Number of Castings. Small series to large production runs, depending on the complexity and/or machinability of the workpiece concerned.

Tolerances. From ±0.4 to ±0.7% of nominal dimensions.

Full Mold Casting (Evaporative Pattern Casting, Lost Foam Casting). The mold is lost (used only once), mold material is usually self-curing.

Pattern. Expendable, foam material.

Process Characteristics (Fig. 7.14). One-piece pattern made of foam material (polystyrene). Shape and dimensions match the part to be cast (taking into ac-

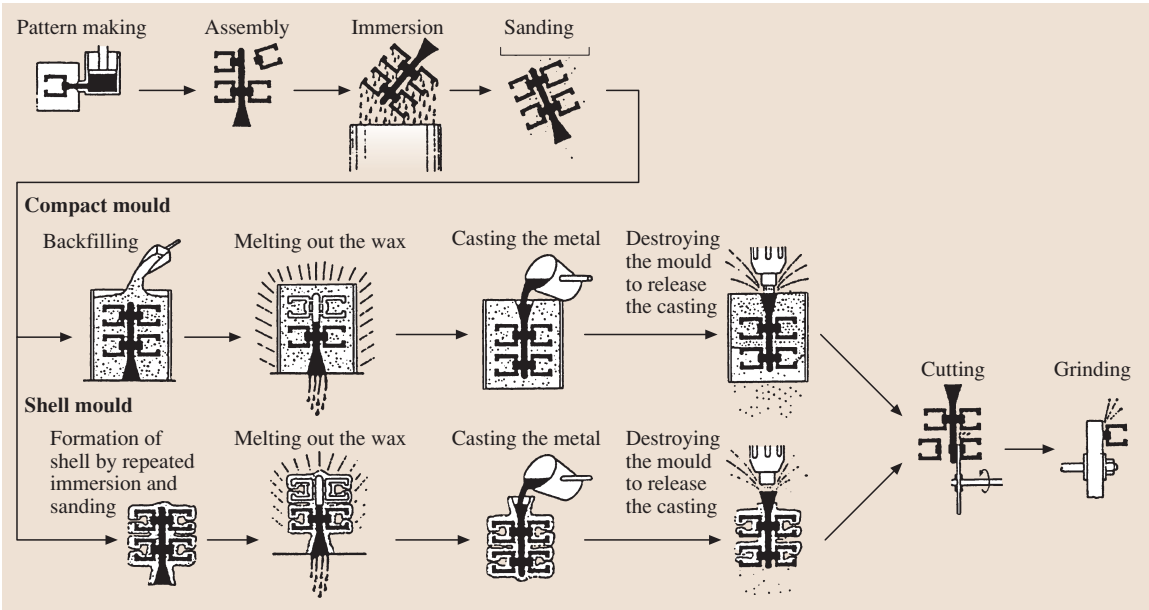


Fig. 7.13 Schematic manufacturing stages in investment casting (after [7.6])

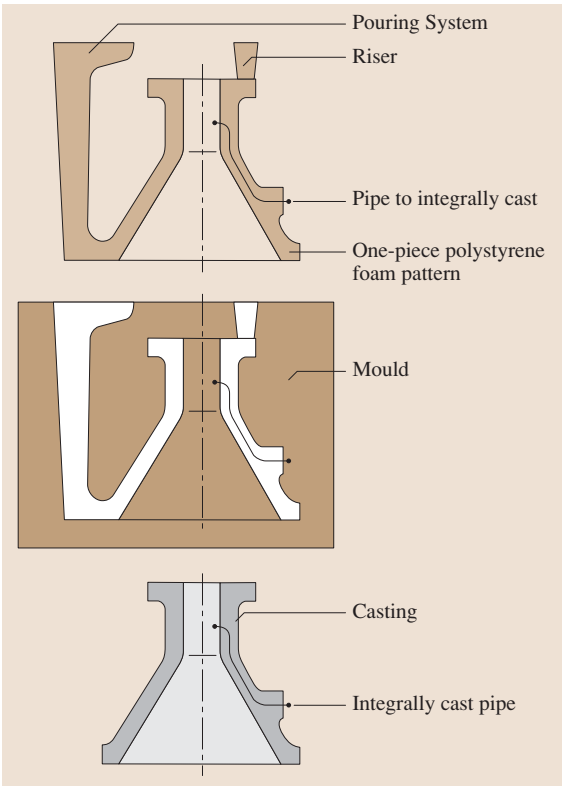


Fig. 7.14 Full mold casting (after [7.5])

count allowance for contraction). The pattern need not be removed from the mold after mold-making. The heat of the molten metal flowing into the full mold vaporizes the pattern, which is continuously replaced by cast metal. Mold joints and cores are usually unnecessary. Bolts, sleeves, lubrication lines, etc. can be integrally cast in. The absence of mold tapers reduces the weight of the casting. The time and cost of manufacture are only a fraction of those encountered with a wooden pattern.

Casting Materials. All metals and alloys that are castable with the current technology, especially those with high casting temperatures.

Weight of Castings. From ≈ 50 kg up to maximum transportable weight; especially suitable for large components.

Number of Castings. Single pieces, small production turns.

Tolerances. From about 3 to 5%.

Magnet Molding. The mold is expendable (used once only), iron granules.

Pattern. Expendable, foam material.

Process Characteristics. Magnetic molding is a type of full mold casting. The casting units, prefabricated from foam material (patterns with pouring gates and runner), are coated with a refractory ceramic (similar to shell molds for precision casting). They are then backfilled with pourable iron granules in a mold box. By applying (or switching on) a DC magnetic field, the iron powder becomes rigid and thus supports the casting unit. After casting and solidification of the die metal, the magnetic field is switched off, causing the iron granules to become pourable again. Then the casting is removed. The iron granules can be reused.

Casting Materials. All metals and alloys that are castable with the current technology. As the thermal conductivity of the magnetizable mold material is higher than that of quartz sand, the cooling rate of the castings is higher and leads to a finer metallographic structure. The properties in use are especially improved in the case of steel castings.

Number of Castings. Single items, small production runs.

Tolerances. From about less than 3 to 5%.

Use of Permanent Molds.

Permanent Mold Casting (Gravity Die Casting). The mold is a permanent mold from cast iron or steel, cores made of steel.

Pattern. None required.

Process Characteristics (Fig. 7.15). Casting takes place by gravity in permanent metal molds. These molds are made in two or more parts for removal

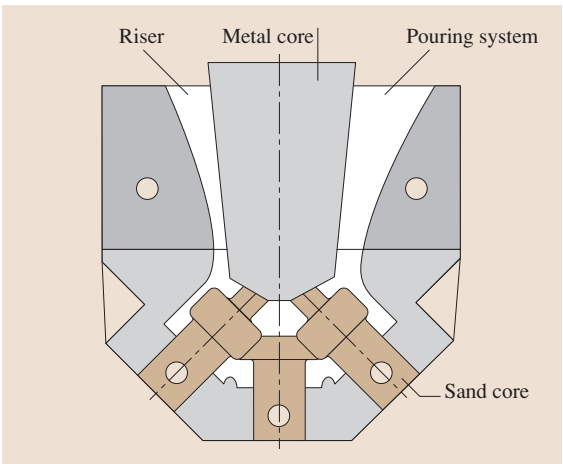


Fig. 7.15 Permanent mold casting (composite mold with metallic and sand cores, latter for undercuts) (after [7.5])

of the finished casting. The higher thermal conductivity of the metal mold compared with molding sand brings about faster cooling of the solidifying molten metal. The result is a relatively fine-grained, dense structure with better mechanical properties than parts made by sand casting. High dimensional accuracy, excellent surface quality, and good reproduction of contours characterize chill casting. This process fully meets the specifications for gas- and liquid-tight valves, owing to the production of a dense structure. A rapid, efficient casting sequence, with machining generally being unnecessary or requiring only small machining allowances, are further features of this process.

Casting Material. Copper-zinc alloys, copper-aluminum alloys, aluminum alloys, magnesium alloys, high-grade zinc alloys, also copper, copper-chromium alloys, super-eutectic aluminum-silicon alloys, lamellar and nodular graphite cast iron.

Weight of Castings. Nonferrous metals and cast iron up to ≈ 100 kg, more depending on equipment. Cast iron for certain purposes up to ≈ 20 t (= 20 000 kg).

Number of Castings. From about 1000 and multiples thereof, depending on the material being cast (e.g. Al ≈ 100 000 castings).

Tolerances. From about 0.3 to 0.6%.

Low-Pressure Permanent Mold Casting. The mold is a permanent mold from cast iron or steel.

Pattern. No pattern required.

Process Characteristics (Fig. 7.16). Casting is carried out under pressure (usually with compressed air) in permanent metal molds. These molds are made in two or more parts for removal of the finished casting. The higher thermal conductivity of the metal mold compared with molding sand brings about faster cooling of the solidifying molten metal. The result is a relatively fine-grained, dense structure with better mechanical properties than parts made by sand casting. The distinguishing feature is the application of pressure, which dispenses with the need for risers on the casting. High dimensional accuracy, excellent surface quality, and good contour reproduction together with a rapid, efficient casting sequence and considerable machining economies are further features of this process. Gas-tight and liquid-tight valves can be efficiently manufactured owing to the dense structure of the casting.

Casting Materials. Light metal, especially aluminum alloys.

Weight of Castings. Up to 70 kg.

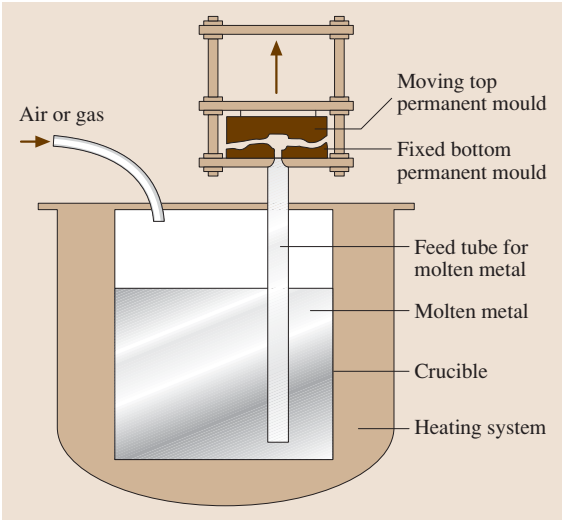


Fig. 7.16 Low-pressure permanent mold casting (after [7.5])

Number of Castings. From about 1000 and multiples thereof.

Tolerances. From about 0.3 to 0.6%.

High-Pressure Die Casting (Die Casting). The mold is a permanent mold, usually from high-tensile hot forming tool steel or special metals.

Pattern. No pattern required.

Process Characteristics (Fig. 7.17). The distinguishing feature of this process is that the molten metal is forced into the two-part permanent mold at high pressure and relatively high speed in pressure die casting machines. Two types of process are distinguished, namely:

- The hot-chamber process. Here the die casting machine and the holding furnace for the molten metal form a unit. The casting assembly is immersed in the molten metal. In each casting operation, a precisely predetermined volume of molten metal is forced into the mold. The hot-chamber die casting process is especially suitable for lead, magnesium, zinc, and tin. The output of components manufactured by this process is considerable, but varies according to the size of the component and the casting material.
- The cold-chamber process. In this process the die casting machine and the holding furnace for the molten metal are separate. After being taken from the furnace, the molten metal is poured into the cold pressure chamber and forced into the mold.

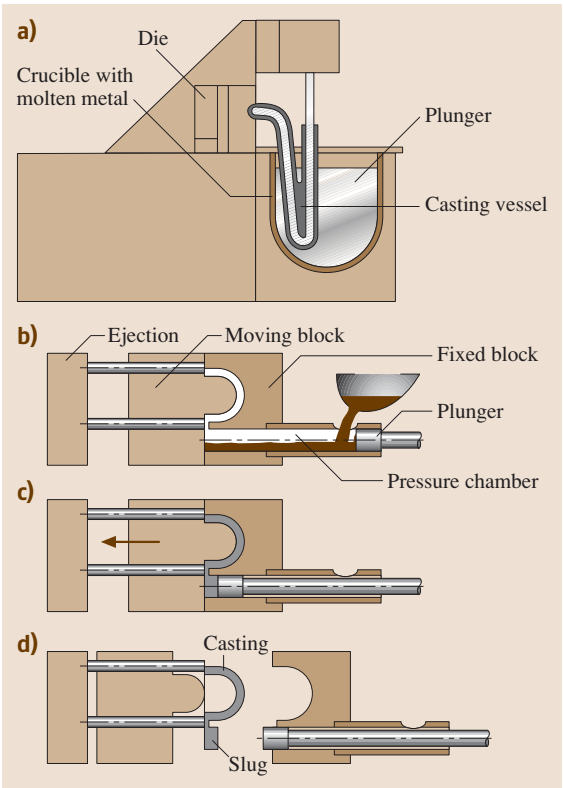


Fig. 7.17a–d High pressure die casting: **(a)** hot chamber process, **(b–d)** cold-chamber process, **(b)** filling of casting chamber, **(c)** plunger for molten metal into die, **(d)** ejection of casting (after [7.5])

The pressure chamber is mounted directly on the runner-side mold block. This process is chiefly suitable for aluminum-based and copper-based alloys, as these would attack the steel casting assembly when molten if the hot-chamber process were employed. Cold-chamber die casting machines do not achieve the rates of output of hot-chamber machines, owing to the nature of the process. Pressure die casting is today one of the most efficient casting processes around. The machines are mostly semiautomatic or fully automatic. Pressure die-cast parts have smooth, clean surfaces and edges. They are extremely dimensionally accurate. Therefore, only the fitting and bearing surfaces, at most, require machining. Very low machining allowances result in short machining times.

Special cases of high pressure die casting are squeeze-casting and thixocasting.

Squeeze-Casting and Thixocasting. In the direct squeeze casting, the die is filled with a defined amount of liquid metal via a trough. The die, which has two or more parts, is closed above by hydraulic pressure, forcing the metal to conform to the die surfaces. The pressure is maintained during solidification. Unlike direct squeeze-casting, the indirect version has proved itself in production and has obtained a foothold in foundries producing castings for special applications. The principle of operation differs from the direct process in that the melt is not poured directly into the die, but into a casting chamber situated below it. The bottom of the chamber is sealed by a piston whose actuating cylinder is fixed to a pivot. After filling, the chamber is swung into position under the die and the piston is raised to fill the die.

Thixocasting takes advantage of the thixotropic properties of metal alloys in the partly solidified state (semisolid state). Thixotropic behavior means that the material behaves as a solid when at rest but flows like a liquid when rapidly deformed, its viscosity falling as the stress increases. To bring this about, the alloy has to be heated to a point in its freezing range between solidus and liquidus temperatures. This point be chosen such that, for example, 40% of the volume is liquid, the rest remaining solid. Aluminum alloys having a long freezing range are suitable for thixocasting. To attain a good thixotropic state, the crystals in the solid solution must be equiaxed in order to ensure the liquid and solid phases flow uniformly and do not separate. Such a structure can be created by so-called *rheocasting*, whereby the melt is mechanically or electromagnetically stirred during solidification at continuous casting. This breaks off or melts off the dendrite arms, which condense to globular shapes when held just above the solidus temperature, resulting in indigenous growth of free-floating crystals with the desired equiaxed structure.

Current practice in thixocasting is as follows: The rheocasted bar is first cutted off in pieces having sufficient weight for the part to be casted. The pieces are then preheated to a predetermined temperature in their melting range. This is done automatically using temperature sensors to attain a definite proportion of liquid metal, often between 35 and 50%. It is placed in the casting chamber of the casting machine and is injected into the mold cavity by the casting piston.

Casting Materials. Materials suitable for pressure die casting are copper-zinc alloys, copper-aluminum alloys, aluminum alloys, magnesium alloys, lead alloys, tin alloys, high-grade zinc alloys. For the hot-chamber

process: lead, magnesium, zinc, and tin alloys. For the cold-chamber process: particularly aluminum- and copper-based materials.

Weight of Castings. Up to 45 kg for light alloys, up to 20 kg for other materials, depending on the material being cast and the working dimensions of the die casting machines.

Number of Castings. Varies widely depending on the material being cast (e.g. Zn alloys ≈ 500 000 castings).

Tolerances. From about 0.1 to 0.4%.

Centrifugal Casting. The mold is a permanent, water-cooled cast iron or steel mold.

Pattern. None required.

Process Characteristics (Fig. 7.18). The centrifugal casting process is used to manufacture hollow products having a rotationally symmetrical hollow space and an axis coinciding with the axis of rotation of the centrifugal casting machine. The external form of the casting is determined by the shape of the mold. The internal form is determined by the effect of the centrifugal force of the rotating mold. The wall thickness of the casting depends on the quantity of molten metal supplied. A variant of the process is centrifugal mold casting, which produces finished hollow or even massive castings using rotating molds. Composite centrifugal casting is also possible, as is centrifugal casting with a flange. The condition on delivery of centrifugal cast Fe, Ni and Co-based alloys is normally (at least) preturned.

Casting Materials. Especially cast iron, cast steel, heavy and light metals.

Weight of Castings. Up to about 5000 kg,

Number of Castings. From about 5000 to over 100 000, depending of the mold material and the casting material. In special cases, e.g. castings of stainless steel and the like, also single pieces and small production runs (from ≈ 104 mm internal diameter upwards).

Tolerances. About 1%.

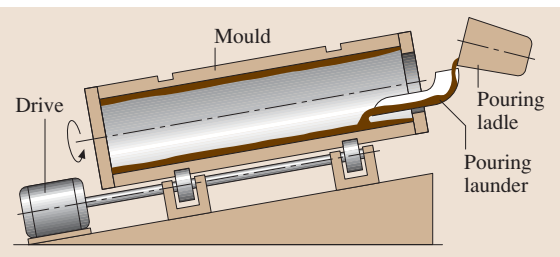


Fig. 7.18 Centrifugal casting (after [7.5])

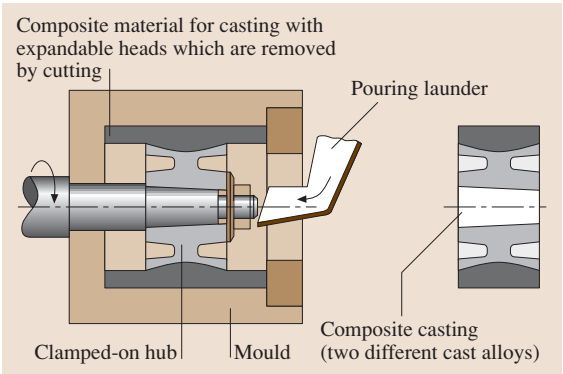


Fig. 7.19 Composite casting (after [7.5])

Composite Casting (Compound Casting). The mold is a lost or permanent mold. Metal mold. e.g. for centrifugal composite casting.

Pattern. None.

Process Characteristics (Fig. 7.19). These types of process are used to cast structural parts from two or more different metallic materials that are firmly joined together. At least one material is poured in the molten state into a mold, which may also be part of a product to be manufactured. For composite casting of various metals and/or alloys in molten or semisolid condition, e.g. in centrifugal casting and for casting in, casting round and lining solid components, which may be made not only of metal but also ceramic. The bonding may be formed by shrinkage, diffusion, or both.

Casting Materials.. All metals and alloys that are castable with the current technology.

Weight of Castings. Up to about 50 kg and over, depending on the manufacturing equipment.

Number of Castings. Medium and large production runs.

Tolerances. From about 0.1 to 0.6%, depending on the process.

7.1.6 Guidelines for Design

Forming by casting enables design ideas to be turned into reality to a particularly high degree, owing to the extensive freedom of design that it offers. A design appropriate for manufacturing, which contributes decisively to the efficient production of a casting, can generally be achieved only by close collaboration between the design engineer and the founder. Forming by casting differs from other forming processes in that the material only receives its shape, material structure, and quality after cooling; with shrinkage – which may

Table 7.14 Contraction of various casting materials (approximate values)

Material	Liquid (max. %)	Solid (max. %)
Lamellar graphite cast iron	3	1
Nodular graphite cast iron	5	2
Cast steel	6	3
Malleable cast iron	5.5	2
Copper alloys	4	2
Aluminum alloys	5	1.25

sometimes be considerable – in the liquid state and during solidification, and appreciable contraction in the solid state (Fig. 7.20, Table 7.14). The contraction in the solid state should be accounted for by means of a suitable allowance (allowance for contraction). The alloying specifications often suffer considerable deviations, owing to obstruction of contraction by ribs, projections, more or less flexible cores, and mold parts (Table 7.15). Provided that they fall within the acceptable dimensional variations or are compensated for by machining allowances, as is usually the case with small castings, they do not present a problem. With large castings, though, empirical values for the deviations due to contraction have to be taken into account when making the pattern. If there is a one-sided obstruction of contraction, e.g. due to the mold or even the shape, especially in the case of longer castings (different cross-sections along their length and consequently different cooling rates and thermal stresses), the castings would distort unless the pattern is curved in the opposite direction. Large wheel centers, for instance, are not infrequently split, e.g. to prevent unacceptable out-

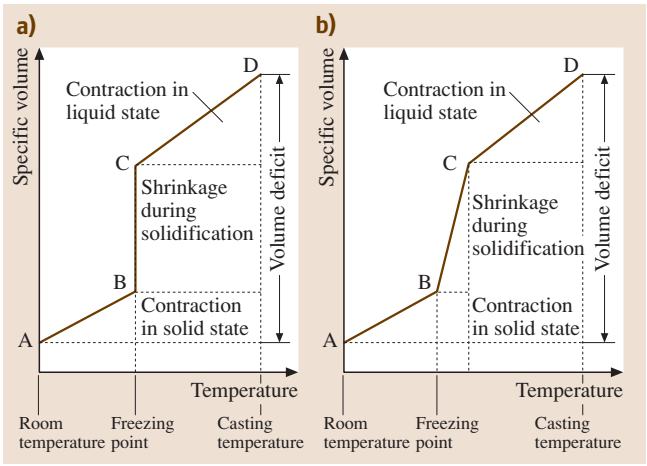


Fig. 7.20a,b Schematic of contraction of metallic materials during cooling from the molten state: **(a)** for pure metals and eutectic alloys, **(b)** for noneutectic alloys (after [7.7])

Table 7.15 Guide values for linear contraction and possible deviations

Casting material	Guide value (%)	Possible deviation (%)
Lamellar graphite cast iron	1.0	0.5–1.3
Nodular graphite cast iron, unannealed	1.2	0.8–2.0
Nodular graphite cast iron, annealed	0.5	0.0–0.8
Cast steel	2.0	1.5–2.5
Austenitic manganese steel	2.3	2.3–2.8
White malleable cast iron	1.6	1.0–2.0
Black malleable cast iron	0.5	0.0–1.5
Aluminum casting alloys	1.2	0.8–1.5
Magnesium casting alloys	1.2	1.0–1.5
Casting copper (electrolytic)	1.9	1.5–2.1
CuSn casting alloys (cast bronzes)	1.5	0.8–2.0
CuSn-Zn casting alloys (gunmetal)	1.3	0.8–1.6
CuZn casting alloys (cast brass)	1.2	0.8–1.8
CuZn (Mn, Fe, Al) casting alloys (special cast brasses)	2.0	1.8–2.3
CuAl (Ni, Fe, Mn) casting alloys (cast aluminum bronzes and cast multicomponent aluminum bronzes)	2.1	1.9–2.3
Zinc casting alloys	1.3	1.1–1.5
Babbitt (Pb, Sn)	0.5	0.4–0.6

of-roundness. Thermal stresses that are not reduced by plastic deformation may, besides distortion, also result in undesirable *relief by cracking*. Therefore, if insufficient attention is paid to contraction of the material being cast at the design stage, taking into account the possibilities afforded by gating of molds and risering, pipes (shrinkage cavities), shrinkage voids, sinks, hot (pipe) cracks, distortion, and stress cracks may form.

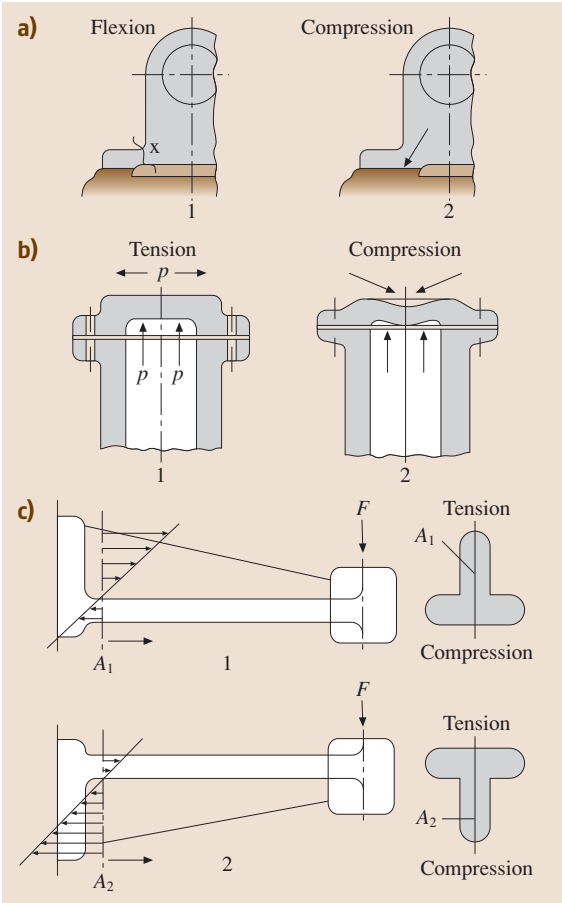


Fig. 7.21a–c Examples for stress-orientated casting design for a material having compression stress strength than tensile strength according to [7.8]: **(a)** pedestal (1 flexurally stressed – inadequate bearing surface, 2 compressively stressed – bearing surface widened); **(b)** cylinder cover (1 tensionally stressed – poor design, 2 compressively stressed – bearing surface – good design); **(c)** wall bracket arm (1 poor cross-sectional arrangement, 2 stress-absorbing cross-sectional arrangement) (after [7.8])

Manufacture-Orientated Design

The correct design of changes in wall thickness with a view to shrinkage during solidification and contraction in the solid state is outlined below.

- Wall thickness graduations should permit directional solidification.
- Junctions formed by the meeting of two or more walls. They should be separated as far as possible, or designed for efficient casting by narrowing the cross-section. Concentrations of material, especially at locations that are inaccessible to feeding, lead to piping.
- Sudden changes in wall thickness should be avoided, as they produce high thermal stresses due to different cooling rates. In addition, there is often increased obstruction of contraction by the mold. The risk of formation of hot cracks (*pipe cracks* between liquidus and solidus temperatures) and stress cracks (during further cooling in the solid state) is therefore high. Locations prone to cracking can be protected by ribs.
- Sharp corners additionally cause a heat build-up (hot sand effect) and, accordingly, not only hot cracks but also porosity due to contraction as well as drawholes.

For a summary of design recommendations see Fig. 7.22.

Stress-Orientated Design

When designing castings, the main stresses occurring during manufacture have to be taken as a basis. Here, the freedom of design offers excellent adaptation to the technical requirements. Forming by casting permits the efficient manufacture of parts of the most complicated kind with high strength in relation to shape. The stress condition of the design can often be made more favorable by suitable ribbing or slight modification (Figs. 7.21 and 7.22).

It is important to know the load-bearing capacity of the materials to be cast. For approximate values for lamellar graphite cast iron see Fig. 7.23.

Examples. A casting made from EN-GJL-150 gray cast iron (top horizontal gray bar) with a wall thickness of 10 mm or a test bar diameter of 20 mm (vertical line) has a tensile strength of $\approx 220\,\mathrm{N/mm^2}$, a hardness of $\approx 220\,\mathrm{HB}$, and a modulus of elasticity of $10\,000\,\mathrm{dN/mm^2}$. For a wall thickness of 45 mm, on the other hand, the tensile strength is $\approx 100\,\mathrm{N/mm^2}$, the hardness is $\approx 130\,\mathrm{HB}$, and the modulus of elasticity almost $8000\,\mathrm{dN/mm^2}$.

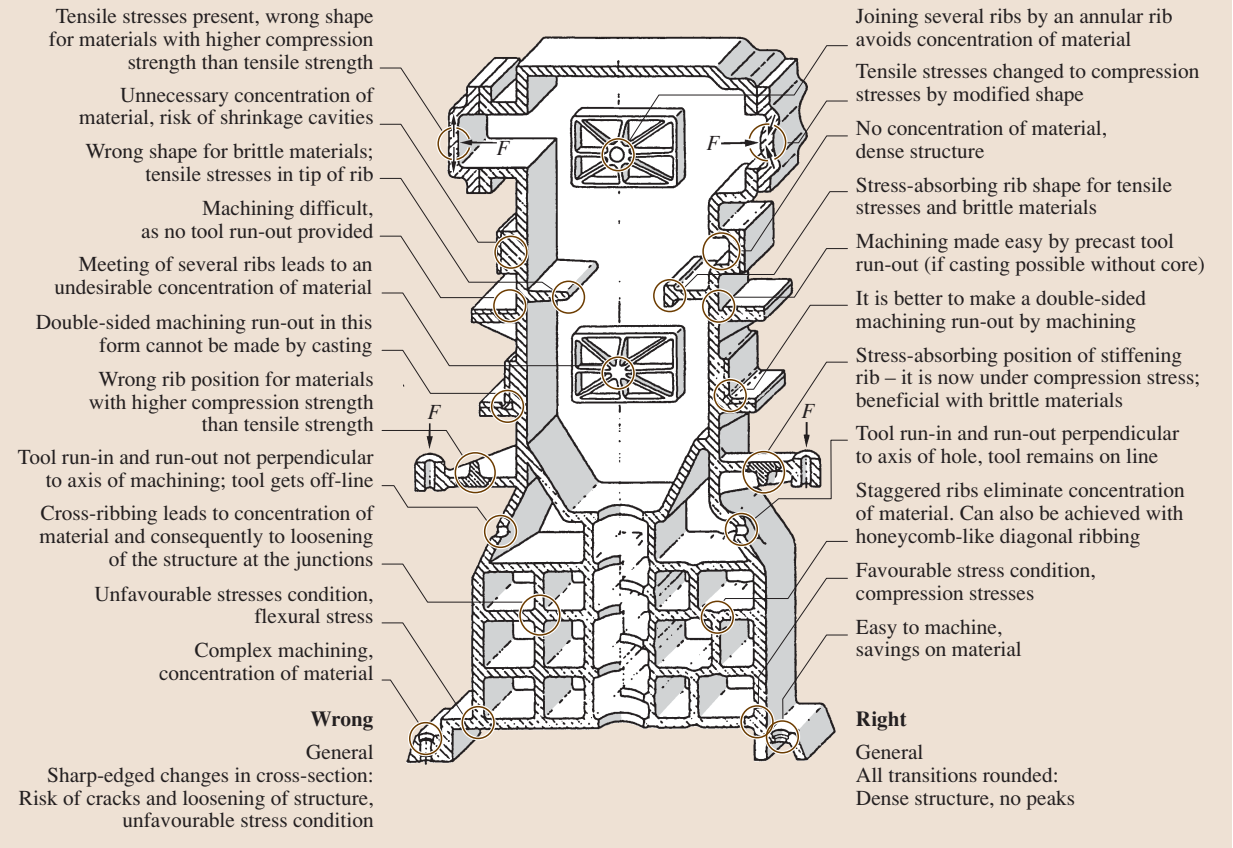


Fig. 7.22 Illustration of important design guidelines (after [7.9])

If however, the tensile strength of this 45 mm thick wall is 220 N/mm^2, a hardness ≈ 180 HB and an modulus of elasticity of ≈ 11 500 dN/mm^2 should be expected. The material grade EN-GJL-30 should be selected. For a wall thickness of 10 mm, this cast iron has a tensile strength of ≈ 350 N/mm^2, a hardness of ≈ 260 HB, and a modulus of elasticity of ≈ 13 000 dN/mm^2.

Near-Net-Shape Manufacturing and Integral Castings

The main advantage of shaping by casting is the realization of near-net-shape production of castings, thereby minimizing cutting processing and drastically shortening the process chains due to fewer process stages. The process chain is dominated up to the finished part by chip-arm shaping.

Development in shaping by casting is focused on two directions. First, the components become increasingly closer to the finished parts. Second, many single parts are aggregated to one casting (integral casting, one-piece-casting). Both directions of development are realized in all variants of casting technology.

For evaluation, the manufacturing examples in Figs. 7.24 and 7.25 were considered from melting up to a commensurable part.

Figure 7.24 shows a technical drawing of a flat part that had previously been produced by cutting stating from a bar, that is now made as a casting (malleable cast iron) using the sand-molding principle. In cutting from a semifinished material, material is utilized at only 25.5%. As a result of shaping by casting, utilization of material was increased to 40%. The effects of shaping by casting become evident in the energy balance (primary energy, cumulative energy demand). For cutting the flat part from a semifinished product, 49 362 GJ/t parts are required. For shaping by casting, 17 462 GJ/t parts are required. Consequently, 64.6% of the energy

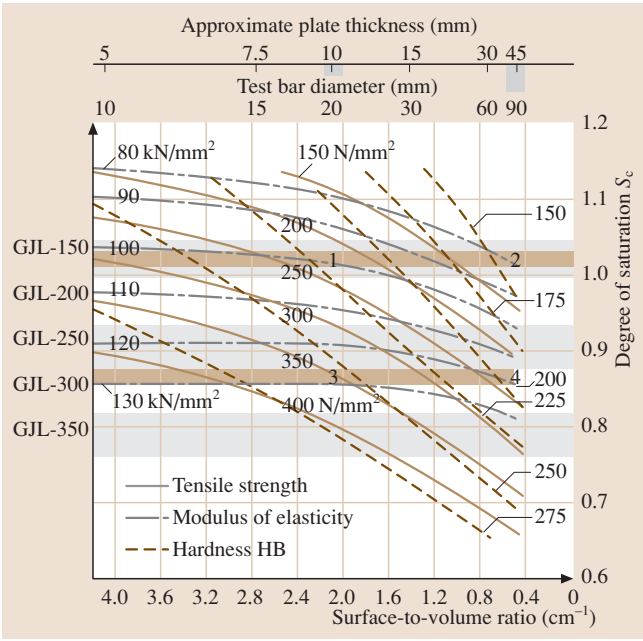


Fig. 7.23 Chart illustrating mechanical properties of gray cast iron. Relationship between chemical composition, rate of cooling, and mechanical properties (tensile strength, hardness, modulus of elasticity) in the casting (wall thickness) and the separately cast test bar [7.10, 11]. Each point on the diagram signifies a specific combination of mechanical properties for a specific material. It also determines the material grade to be selected◀

can be saved. Compared to cutting of semifinished steel material, for part manufacturing about a third as much primary energy is required.

The doorway structure of an Airbus passenger door (PAX door: height about 2100 mm; width about 1200 mm) is illustrated in Fig. 7.25.

The conventional manufacturing of the doorway structure as practiced until now, apart from the standard parts such as rivets, rings, and pegs, 64 milling parts were cut from semifinished aluminum materials with very low material utilization. Afterwards, those parts were joined by about 500 rivets.

As an alternative technological variant, it is proposed that the doorway structure be made of three cast segments. Assuming almost the same mass, in production from semifinished materials, the ratio of chips amounted to about 63 kg, whereas in casting, this can be reduced to about 0.7 kg. Thus, in casting, the chip ratio amounts to only 1% in comparison to the present manufacturing strategy. In the method start-

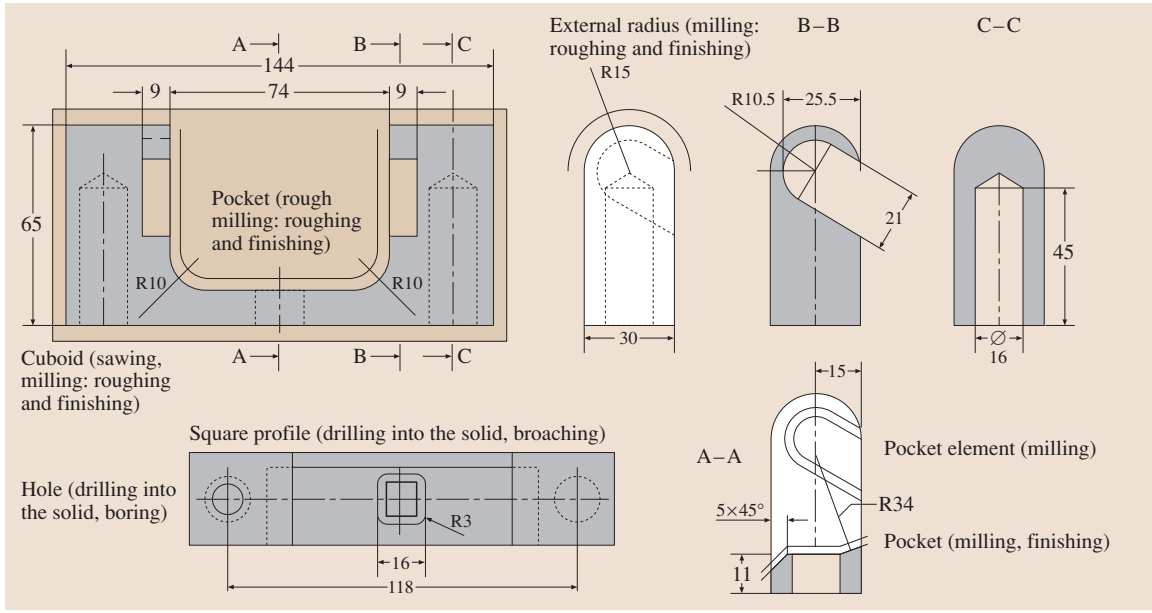


Fig. 7.24 Example: Flat part (after [7.12])

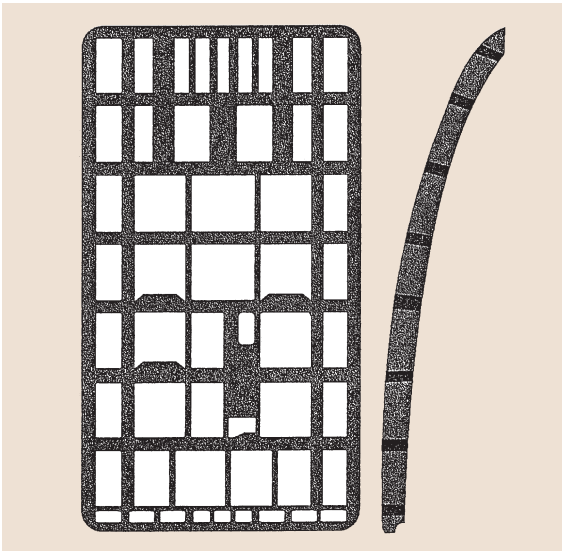


Fig. 7.25 Example: Airbus doorway (after [7.12])

ing from the semifinished material, about 175 kg of material have to be molten; however, in shaping by casting, this value is about 78 kg – that is 44.6%. As a result of the energy balance (primary energy, cumulative energy demand), about 34 483 MJ are required for manufacturing the doorway from the semifinished material. However, in shaping by casting, 15 002 MJ are needed – that is about 46%. The result of having drastically diminished the cutting volume due to near-net shaping can be clearly proven in the energy balance: in the variant starting from the semifinished material, 173 MJ were consumed for cutting, in casting, less than 2 MJ.

In contrast to the studies mentioned above, today, the Airbus door constructions that have been cast as one part only are used. 64 parts are aggregated to one casting (integral casting).

7.1.7 Preparatory and Finishing Operations

Melting of Materials for Casting

For transforming the metal to be cast and the additives into the molten state, a wide variety of melting equipment – e.g. shaft (cupola), crucible, and hearth-type furnaces – is available. These furnaces are heated with coke, gas, oil, or electricity. The most important types of melting equipment are: for cast iron, including malleable cast iron: cupola (shaft) furnaces, induction furnaces, rotary kilns (oil-fired); cast steel: electric arc furnaces, induction furnaces; nonferrous metal castings: induction furnaces, electrically, gas-heated or oil-heated crucible furnaces.

Cleaning of Castings

The molds are emptied by means of emptying jiggers. The sand adhering to the casting is generally removed by means of abrasive blasting equipment employing, without exception, steel shot or steel grit made from wire.

Heat Treatment

Many materials only obtain the physical and technological characteristics required in use from heat treatment. This treatment requires the use of electrically heated or gas- or oil-fired furnaces in continuous or batch operation. Their size is matched to the size and quantity of the castings and their mode of operation to the wide variety of heat treatment processes.

Inspection and Testing Methods

The diverse demands made on the casting, which become greater with advances in technology, and the trend towards lightweight construction and thus more efficient use of materials inevitably lead to stringent requirements with regard to casting quality, with particular emphasis on consistency. Inspections of the process and the castings begin with checking of the metallic and nonmetallic feedstocks and end with the final inspection of the castings. Materials and workpieces are mainly tested by means of nondestructive testing methods such as radiographic (EN 12681), ultrasonic (EN 12680, part 1, 2 and 3), magnetic powder (EN 1369), and liquid penetrant (EN 1371 part 1 and 2) testing [7.13–21]. Destructive tests, e.g. tensile, notched-bar impact and bending, are usually carried out with specimens cast either separately or as an appendage to the casting; in exceptional cases, specimens taken from the casting itself may be used.

In the last 20 years the certification of quality management systems on the basis of ISO 9000 etc. and QS 9000 has risen appreciably in the foundry industry.

7.2 Metal Forming

7.2.1 Introduction

Metal-forming is the manufacturing through plastic (permanent) change of the form of a solid body by preserving both the mass and the cohesion. The term forming should be used for controlled plastic straining with a predefined target shape, whereas the term deforming should be used for unwanted or uncontrolled plastic straining. Basic advantages of metal-forming processes as compared to alternative processes such as casting and machining are [7.22, 23]:

- High material utilization and hence high energy conservation
- High productivity with short production times
- High dimensional and shape accuracy within certain tolerances
- Superior mechanical material properties of the product (especially for dynamic loadings)

On the other hand, these processes are exposed to the following disadvantages:

- Due to the high loads required for plastic forming, the tools and machines are expensive requiring minimum batch sizes for economic production.
- The limited formability of metals restricts the range of product geometries.
- The process usually requires a high level of engineering including analytical modeling, numerical analysis (process simulation), and also extended experience.

Metal-forming processes can be classified according to various criteria. An academic classification is done according to the dominant stress state existing in the deformation zone. Accordingly forming processes can be classified into compressive, tensile, tensile–compressive, bending and shearing processes. A more practical classification is according to the type of product by which two classes of processes are identified: bulk forming and sheet forming processes. In bulk forming processes the workpieces have spatial geometries (i. e. their geometries are more or less balanced in all space-directions). During forming large changes in the cross-sections and in the thickness of the products are found. The material flows in all directions. Generally, multiaxial compressive stress states exist in the deformation zone. Larger relative forming forces are needed. Bulk forming can be done either with workpieces at room temperature (cold forming) or at an elevated temperature (warm or hot forming). Where hot forming is done with workpieces heated over their recrystallization temperature, warm forming is performed with workpieces between the room and the recrystallization temperature or slightly over the recrystallization temperature. Examples of typical bulk forming processes are given in the following.

Figure 7.26 shows the primary forming process of flat rolling: an initial blank is reduced in thickness by means of two rolls. The width of the workpiece remains usually approximately constant. The product is used as initial workpiece for other basically sheet forming processes. By rolling also the casting-microstructure is changed to a more homogenous defect-free microstructure.

Figure 7.27 shows the basic process of wire drawing. Here, an initial rod or wire is forced through a conical die in order to reduce its diameter. Due to the application of a drawing force the area reduction is limited to about 20%. Extrusion is demonstrated in Fig. 7.28.

Similar to wire drawing the cross-section of a rod is reduced by forcing it through a die opening. However this time the workpiece is pushed into the die instead of pulled through. The extrusion process shown in the figure produces continuous profiles with various cross-sections. These are usually semifinished products and are processed further in subsequent processes. Whereas

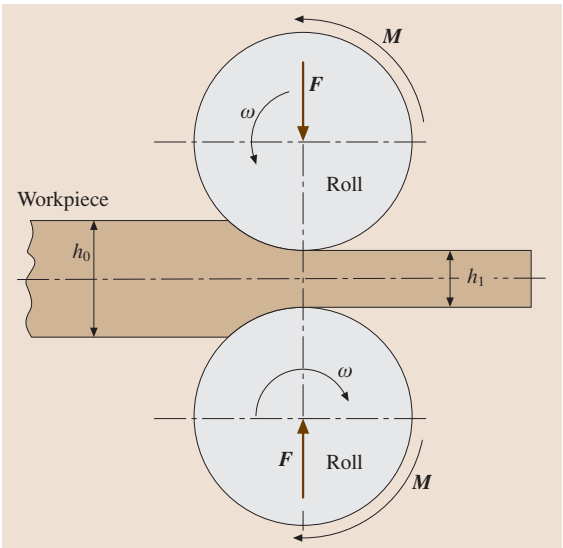


Fig. 7.26 Flat rolling

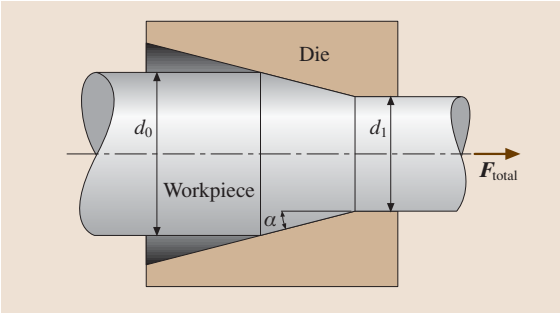


Fig. 7.27 Wire drawing

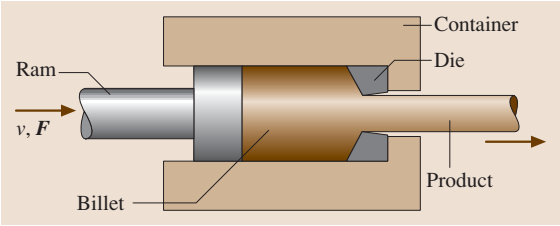


Fig. 7.28 Extrusion

the extrusion process in Fig. 7.29, also called impact extrusion or cold-forging, aims to produce discrete parts, here, net-shaped or near-net shaped parts, i. e. parts that need no or only minor machining or similar shaping operations are formed.

Figure 7.30 shows typical forging processes. In these processes the initial – usually heated – workpiece is formed between two dies. If the dies are flat this process is called upsetting or open-die forging (Fig. 7.30a). If the part is formed between shaped dies and during forging an excessive material flow occurs in form of a flash, this process is called impression die forging (Fig. 7.30b). Closed die forging is the forging process in which no flash is produced. This is also called trapped-die forging.

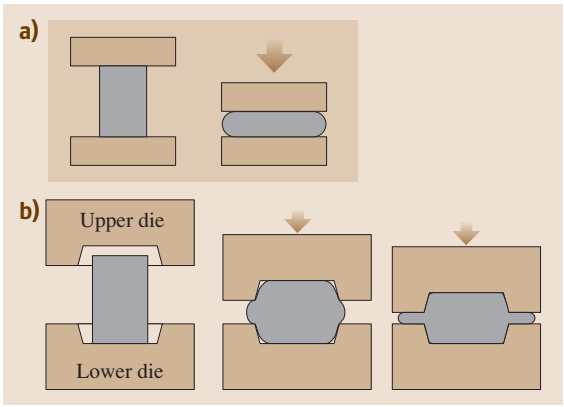


Fig. 7.30a,b Forging: (a) upsetting, (b) impression forging

Two further bulk forming processes that have sheet-like products are shown in Fig. 7.31. Ironing (Fig. 7.31a) is the process in which the wall-thickness of a cup is reduced. This is considered as a bulk forming process since the thickness is reduced and the stress states resembles a typical bulk forming process. Figure 7.31b shows a flow turning process. Also here the thickness of a hollow workpiece is reduced and the stress state is typical for bulk forming.

Sheet forming processes, on the other hand, deal with planar workpieces (sheets, plates) and aim to produce hollow pieces with almost constant wall thickness. Here a change of wall thickness is not intended. Generally two-axial stress states exist. These are either tensile–compressive or tensile–tensile, and the third stress component normal to the sheet plane is zero or nearly zero. Sheet forming processes are usually conducted with workpieces that are not heated. Figures 7.32–7.34 show some examples of typical sheet forming processes. Figure 7.32 describes the bending of sheets. The mechanics of deformation during bending depends

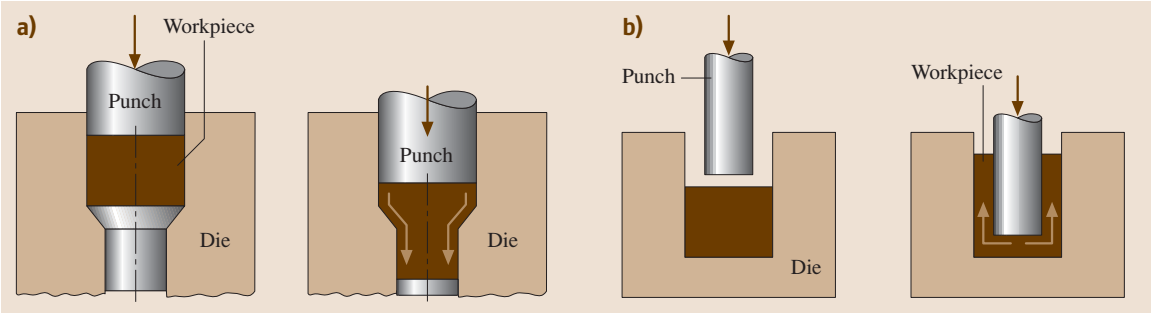


Fig. 7.29a,b Impact extrusion: (a) forward rod extrusion, (b) backward can extrusion

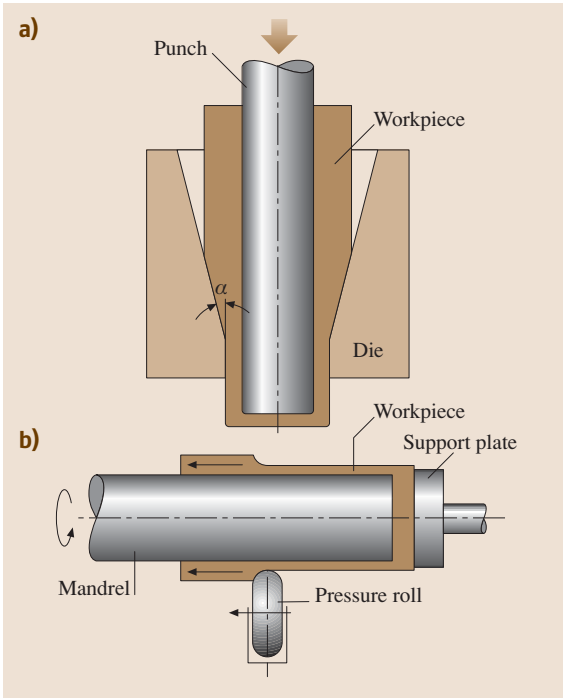


Fig. 7.31a,b Ironing and flow turning: (a) ironing, (b) flow turning

very much on the ratio of the sheet thickness to the bending radius. Mostly the thickness of the sheet varies only slightly during bending. In the figure the free bending process is shown. Bending can be also conducted in dies.

The process of stretching (Fig. 7.104) consists of applying tensile forces in the plane of the sheet. The thickness of the sheet is reduced considerably in stretching. Figure 7.105 shows the basic process of deep-drawing. During this process an originally mainly flat sheet is drawn by means of a punch into a die producing a hollow cup. The basic deformation occurs in the flange of the workpiece and the stress state here consists of a tensile and a compressive component. The sheet is thickening at the outer rim. In this process

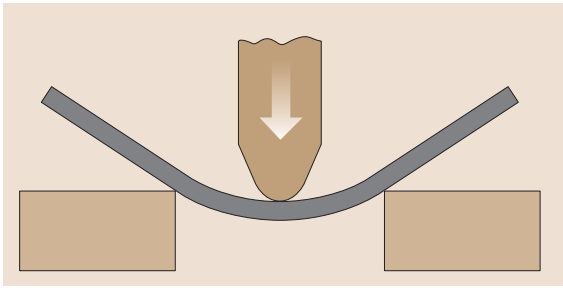


Fig. 7.32 Bending

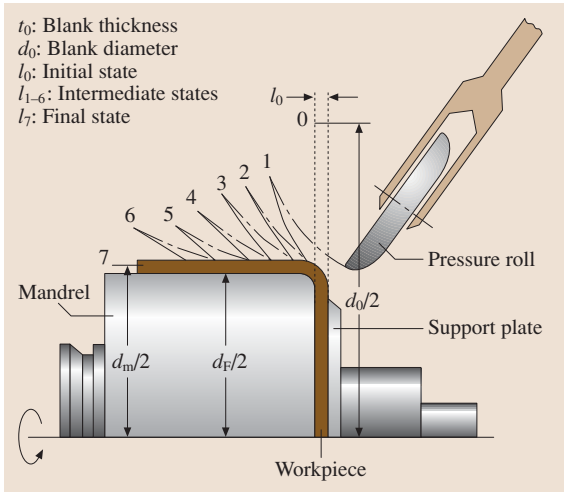


Fig. 7.33 Spinning

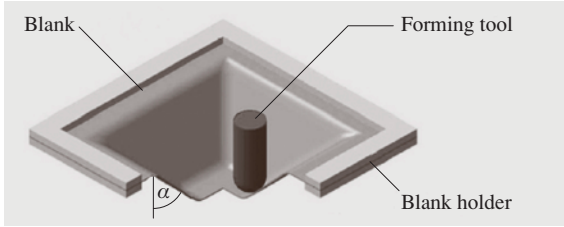


Fig. 7.34 Incremental sheet forming

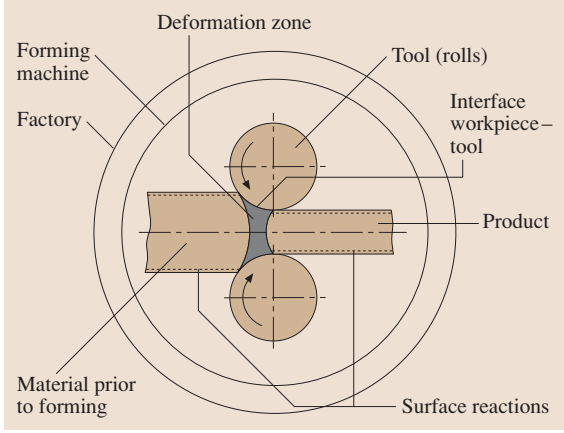


Fig. 7.35 Metal forming system (after [7.22, 23])

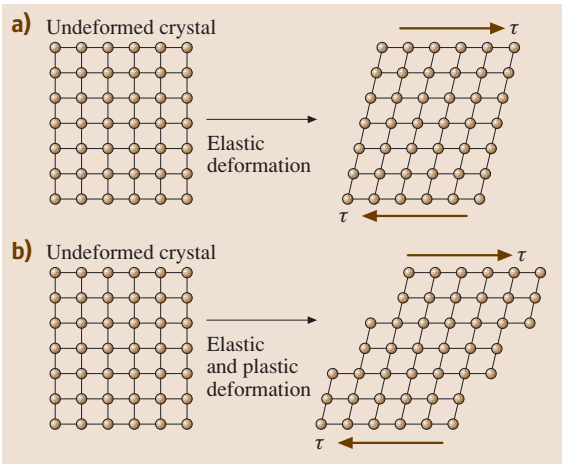


Fig. 7.36a,b Elastic and plastic deformations of the lattice structure

also bending at the die radius and stretching in the cup walls occur.

Figure 7.33 exhibits the process of spinning. Spinning is a typical incremental process in which an originally flat sheet is formed by means of a roll towards a mandrel where the sheet is rotating and the roll is performing a translatory motion.

Another recent incremental sheet forming process is sketched in Fig. 7.34. In this process a punch controllable in three to five axes is used to shape an originally flat sheet incrementally. The process is much slower than the classical deep-drawing or stretching process and has the advantage of being flexible. Besides solid tools also pressure media can be used in sheet forming operations (Figs. 7.114–7.116). Hydroforming, aqua-drawing, and tube-hydro-forming are various process variants of this process family.

Analysis of metal forming processes can be done most efficiently in a system approach. Basic items of the metal forming system are shown in Fig. 7.35 for the flat rolling process. The first item is the material prior to forming. The deformation zone is the second item. Material properties of the plastically formed workpiece, the product, constitute the third item. The tools, here the rolls, make up the another item. The interface between the workpiece and the tools is item five. The surface reactions of the workpiece before and after deformation constitute the sixth item of the system of metal forming. The two final items are the forming machine and the factory. All items influence each other and must, therefore, be mutually taken into considerations.

7.2.2 Metallurgical Fundamentals

Mechanisms of Plastic Deformation

If a metal crystal is exposed to shear stresses, first the lattice will deform elastically as shown in Fig. 7.36a. By removal of the shear stress the lattice will recover its original form. If the shear stress is increased the crystal will deform elastically and plastically, such that upon removal of stresses the crystal will not recover completely its original shape (Fig. 7.36b). Plastic deformations occur by the movement of atomic layers. This movement occurs on so-called slip planes along slip directions [7.24].

The theoretical shear stress necessary to move one part of the lattice along a slip plane with respect to the other part is estimated as

$$\tau_{\text{theoretical}} = \frac{G}{30} \text{ to } \frac{G}{2\pi}\,, \tag{7.1}$$

where G is the shear modulus of the metal. For typical metals the theoretical shear strength as calculated by (7.1) is tabulated in Table 7.16.

A second mechanism of plastic deformation is twinning (Fig. 7.37). During twinning atoms on one side of the twinning plane are almost instantaneously moved to a mirror position of the other side of the plane. Twinning is a preferred mechanism of plastic deformation in hexagonal closed packed (HCP) metals such as titanium, magnesium and zinc alloys. Also, for other metals twinning is preferred if

Table 7.16 Theoretical and actual shear strength for typical metals

Metal	Shear modulus (MPa)	Theoretical shear strength (MPa)	Actual shear strength (MPa)
Steel	75 800	2527–12 063	150–750
Aluminum alloys	27 500	917–4377	50–150
Copper alloys	41 400	1380–6589	100–250
Titanium alloys	44 800	1493–7130	350–800

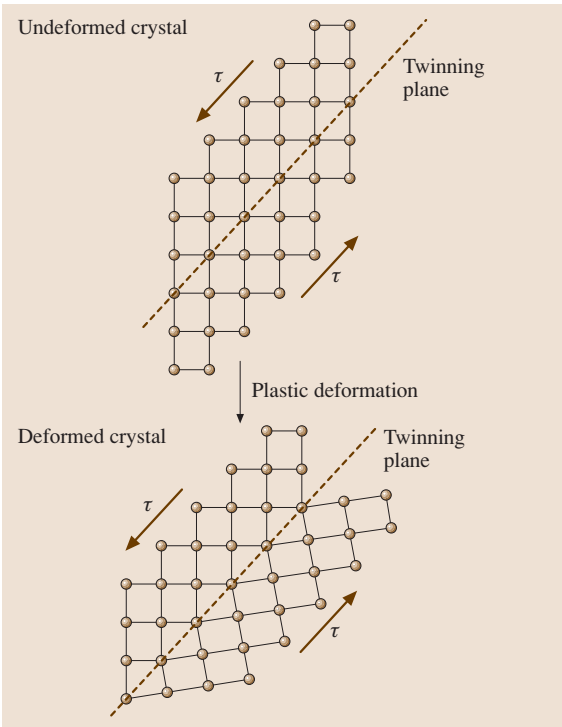


Fig. 7.37 Twinning mechanism for plastic deformation

the strain rates are very high, since plastic deformation by twinning needs much less time than by sliding.

Actual shear strengths of metals are up to several orders of magnitudes less than the theoretical ones (Table 7.16). Real metals show crystal defects called dislocations that facilitate another mechanism of slip that requires much less shear stress than the slip in a defect-free crystalline structure. Figure 7.38 shows that the dislocation shear occurs only incrementally over the shear plane and hence requires much less shear stress. Dislocations are virtually present in any metal introduced during solidification. In annealed metals the length of dislocation lines are about 10^4 to 10^8 cm/cm^3 of metal. As plastic deformation proceeds new dislocations are generated through various mechanisms such as the Frank–Read mechanism. As the number of dislocations increase, they start to hinder each other's motion so that an increased shear stress is necessary to move them. This increase in yield strength is called strain hardening. In cold worked metals the density of dislocations rises up to 10^{10} to 10^{12} cm length per one cm^3 of metal. During plastic deformation about 85 to 90% of the deformation energy is dissipated as heat. Recent measurements indicate that this ratio varies with the amount of plastic strain. The remainder of the energy is stored in the lattice as strain energy and is directly related to the dislocations.

Other mechanisms increasing the yield strength are the solid solution hardening, particle (dispersion) hardening, and grain size hardening. These mechanisms can be superposed to supply an analytic relation for the yield strength as

$$Y = Y_0 + \Delta Y_\mathrm{s} + \Delta Y_\mathrm{d} + \Delta Y_\mathrm{p} \,, \tag{7.2}$$

where Y_0 is the yield stress in an pure single crystal metal with some dislocations, ΔY_s is the increase in yield strength due to solid solution hardening, ΔY_d is the increase due to dispersion hardening, and ΔY_p is the increase due to phase boundaries. For ferritic steel, for instance, Y_0 is about 30 MPa, and

$$\Delta Y_\mathrm{s} = \sum_i k_i x_i \,, \tag{7.3}$$

where x_i is the weight percentage of the alloying element and k_i is a weight factor given in Table 7.17.

Yield strength increase due to grain refinement is given by the Hall–Petch relationship

$$\Delta Y_\mathrm{p} = \frac{k_y}{\sqrt{d}} \,, \tag{7.4}$$

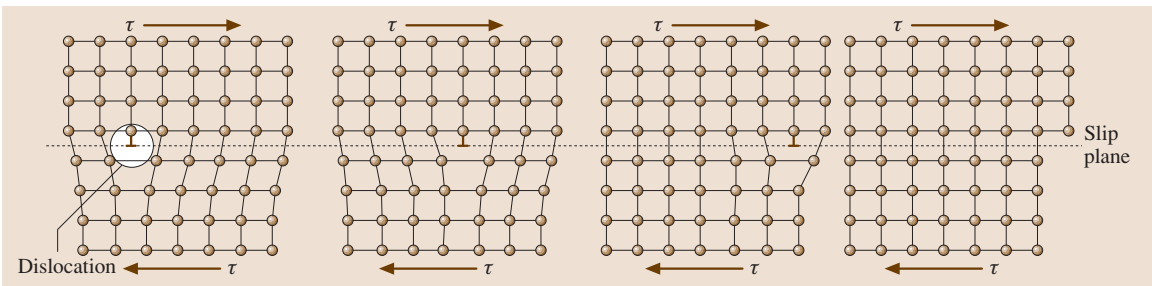


Fig. 7.38 Plastic deformation by dislocation motion

Table 7.17 Weight factors for solid solution hardening (after [7.25])

Element	Si	Mn	P	Ni	Mo	Cu	Sn	Al	N
k_i	81	18	590	8	15	40	130	24	2300

where d is the grain size and the constant k_y ranges between 15 to 24 MPa mm$^{1/2}$ for common steels. This relationship reflects the fact that grain boundaries impose geometrical constraints for the dislocation motion and hence increase the necessary shear stress to move them. The value of ΔY_d depends on the form and morphology of the carbides generated in ferrite.

Reverting Strain Hardening

If a polycrystalline metal is plastically deformed at temperatures below one third of its melting temperature T_m (in K) the grain shapes change (leading to texture) and the dislocation density increases, leading to strain hardening. Strain hardening is naturally accompanied by a reduction of ductility (Fig. 7.39). Although the strain hardened state of the material is mechanically stable it is thermodynamically unstable. Hence, at elevated temperatures this state of the plastically deformed material can be altered.

Recovery. Heating the deformed metal in a range of $0.3T_m < T < 0.5T_m$ will activate diffusion of atoms. This diffusion of atoms enables the motion of some dislocations, which will cancel each other or restructure themselves. Stored energy is relieved. Residual stresses will be removed, the yield stress will reduce slightly, and ductility will increase. Electrical and thermal properties will be recovered as well.

Recrystallization. If the deformed metal is heated at $T > 0.5T_m$ diffusion processes continue such that similar dislocations start to align in certain regions to form low-angle grain boundaries. If the number of aligned dislocations increases they build the large-angle grain boundaries. This process is called recrystallization. The movable collocated dislocations are the recrystallization nuclei. During recrystallization a refined grain structure is built and the yield stress of the metal drops to its original value, and the ductility also increases back to its preworked values (Fig. 7.40).

After recrystallization the new grains show a preferred orientation that is related to the deformation texture and also to the character of the grain boundaries. The temperature at which the whole structure is completely recrystallized within 1 h is called the recrystallization temperature of that metal (Table 7.18). The heat treatment involving recrystallization is called annealing. Recrystallization can only occur after a minimum amount of plastic deformation, since a minimum number of dislocations must be mobilized for forming the large-angle grain boundaries. The more the metal has experienced plastic deformation, the lower is its recrystallization temperature. Time and temperature and prior plastic deformation have interchangeable roles (Fig. 7.41). At high temperatures and following large prior cold plastic deformation, after recrystallization is completed grains may continue to grow such

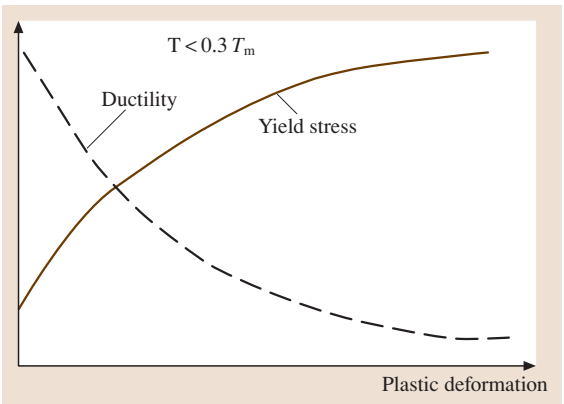


Fig. 7.39 Decrease of ductility and increase of yield stress for temperatures below 0.3× the melting temperature (in K)

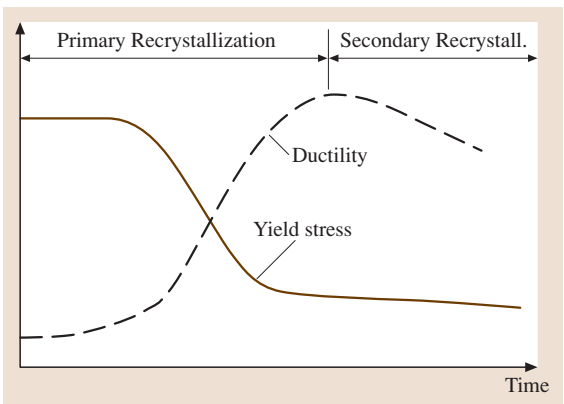


Fig. 7.40 Recrystallization of cold formed metals ($T > 0.5T_m$)

Table 7.18 Recrystallization temperatures for typical metals

Material	Recrystallization temperature (°C)	Material	Recrystallization temperature (°C)
C-Steel	550	Sn	0 to 40
Pure Al	290 to 300	Zn	50 to 100
Dur-Al	360 to 400	Mo	870
Cu	200	W	900 to 1000
Lead	−50 to 50	Ni	400 to 600

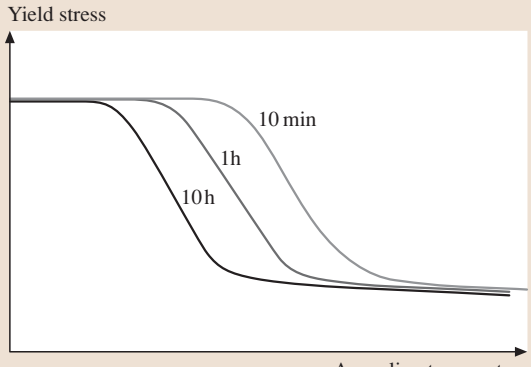


Fig. 7.41 Effect of time and cold forming on recrystallization

that larger grains are built than before recrystallization. This process is called secondary recrystallization and results a material with less superior properties than after the primary recrystallization (such as lower ductility, orange peel effect in deep drawing and bending).

7.2.3 Theoretical Foundations

The theoretical framework of plastic deformation [7.26, 27] is based on four phenomenological columns: The basic concepts of strain and stress, the flow condition, the flow rule, and the flow curve. Utilizing these columns in conjunction with the equilibrium equations analytical and numerical methods are established. Also physical simulation makes use of these fundamentals.

Basic Concepts

The Concept of Strain. Intensity of plastic deformation in metal forming is measured by the true strain increment $d\varepsilon$. Considering the uniaxial tension test (Fig. 7.42) in which a uniform rod with current length ℓ and current cross-sectional area A is exposed to an axial force F, the instantaneous strain increment in axial direction is defined as

$$d\varepsilon = \frac{d\ell}{\ell} . \tag{7.5}$$

Assuming uniform straining the total strain at any point in the rod experienced from the original length ℓ_0 to the final length ℓ_1 is given by

$$\varepsilon = \int_{\ell_0}^{\ell_1} \frac{d\ell}{\ell} = \ln\left(\frac{\ell_1}{\ell_0}\right) . \tag{7.6}$$

Due to the logarithmic expression this strain is also called the logarithmic or natural strain. Compared to the engineering strain the true strain describes finite deformations correctly, is additive between different straining steps, and supplies the same strain values for the same amount of elongations and compressions.

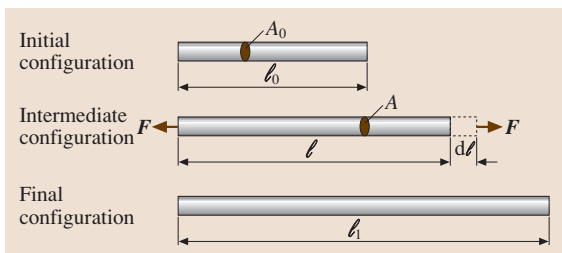


Fig. 7.42 Uniaxial tension of a rod

The total strain given in (7.6) is the sum of elastic (recoverable) and plastic (permanent) strains. However, in metal forming processes elastic strains are several orders of magnitude less than the plastic strains, so the elastic strains can be neglected

$$\varepsilon = \varepsilon^{\text{elastic}} + \varepsilon^{\text{plastic}} \approx \varepsilon^{\text{plastic}} \,. \tag{7.7}$$

Therefore, in the forthcoming sections the superscripts will be omitted.

Strain Rates. If the elongation $\mathrm{d}\ell$ takes place in time of $\mathrm{d}t$, the strain rate can be defined as

$$\dot{\varepsilon} = \frac{\mathrm{d}\varepsilon}{\mathrm{d}t} == \frac{\mathrm{d}\ell/\ell}{\mathrm{d}t} == \frac{\mathrm{d}\ell/\mathrm{d}t}{\mathrm{d}\ell} = \frac{v_{\text{tool}}}{\ell} \,, \tag{7.8}$$

where v_{tool} is the velocity of extension of the rod as imposed by the tool. The concept of strain rates is extended to three-dimensional general deformations by introducing the velocity field for each particle in three Cartesian directions (v_x, v_y, v_z). The components of the strain rate tensor are then defined by

$$\dot{\varepsilon}_{ij} \equiv \begin{pmatrix} \dot{\varepsilon}_{xx} & \dot{\varepsilon}_{xy} & \dot{\varepsilon}_{xz} \\ \dot{\varepsilon}_{yx} & \dot{\varepsilon}_{yy} & \dot{\varepsilon}_{yz} \\ \dot{\varepsilon}_{zx} & \dot{\varepsilon}_{zy} & \dot{\varepsilon}_{zz} \end{pmatrix}$$
$$= \begin{pmatrix} \partial v_x/\partial x & \frac{1}{2}(\partial v_x/\partial y + \partial v_y/\partial x) & \frac{1}{2}(\partial v_x/\partial z + \partial v_z/\partial x) \\ & \partial v_y/\partial y & \frac{1}{2}(\partial v_y/\partial z + \partial v_z/\partial y) \\ \text{symmetric} & & \partial v_z/\partial z \end{pmatrix} . \tag{7.9}$$

Total Strains. For the three-dimensional general deformation state the strain rates can be integrated over the time if and only if all shear strain rates are zero and the straining path, that is, the ratio of the normal strains, is constant over the respective time increment. In this case, the normal strains are also the principal strains. The total strains can then be determined for the deformation depicted in Fig. 7.43

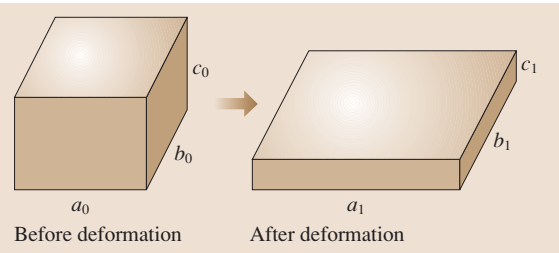


Fig. 7.43 Principal strains

$$\begin{aligned} \varepsilon_1 &= \ln(a_1/a_0) \,, \\ \varepsilon_2 &= \ln(b_1/b_0) \,, \\ \varepsilon_3 &= \ln(c_1/c_0) \,. \end{aligned} \tag{7.10}$$

Obviously, total strains can be used only in idealized cases. Here, the Arabic indices denote the principal directions.

Principle of Volume Constancy. Experimental observations have shown that the volume during plastic deformations remains approximately constant. Referring to Fig. 7.43 this means

$$V = \underbrace{a_0 \times b_0 \times c_0}_{\text{initial}} = \underbrace{a_1 \times b_1 \times c_1}_{\text{final}} = \text{const} \,. \tag{7.11}$$

Or in terms of strain-rates and strains:

$$\begin{aligned} & \dot{\varepsilon}_1 + \dot{\varepsilon}_2 + \dot{\varepsilon}_3 = 0 \\ \text{or} \quad & \varepsilon_1 + \varepsilon_2 + \varepsilon_3 = 0 \\ \text{or} \quad & \varepsilon_x + \varepsilon_y + \varepsilon_z = 0 \,. \end{aligned} \tag{7.12}$$

The Concept of Stress. Referring to Fig. 7.42 the true stress (also called Cauchy stress) in the axial direction is defined by

$$\sigma = \frac{F}{A} \,, \tag{7.13}$$

where A is the current area of the rod. This stress must be carefully distinguished from the engineering stress that refers to the original cross-sectional area.

The three-dimensional generalization of the internal forces according to (7.13) leads to the stress tensor at a point

$$\boldsymbol{\sigma}_{ij} = \begin{pmatrix} \sigma_{xx} & \tau_{xy} & \tau_{xz} \\ \tau_{yx} & \sigma_{yy} & \tau_{yz} \\ \tau_{zx} & \tau_{zy} & \sigma_{zz} \end{pmatrix} \,. \tag{7.14}$$

This tensor is also symmetric like the strain rate tensor. Two important special cases of the general stress tensor are the uniaxial stress state (with only the σ_{xx}-component being nonzero) and the plane stress state (all components with a z-index are zero).

The three independent stress invariants are defined as

$$\begin{aligned} I_1 &= \sigma_{xx} + \sigma_{yy} + \sigma_{zz} = \sigma_1 + \sigma_2 + \sigma_3 \,, \\ I_2 &= \tau_{xy}^2 + \tau_{yz}^2 + \tau_{zx}^2 - \sigma_{xx}\sigma_{yy} - \sigma_{yy}\sigma_{zz} - \sigma_{zz}\sigma_{xx} \\ &= -(\sigma_1\sigma_2 + \sigma_2\sigma_3 + \sigma_3\sigma_1) \,, \end{aligned}$$

$$I_3 = \sigma_{xx}\sigma_{yy}\sigma_{zz} + 2\tau_{xy}\tau_{yz}\tau_{zx} - \sigma_{xx}\tau_{yz}^2 - \sigma_{yy}\tau_{zx}^2 - \sigma_{zz}\tau_{xy}^2 = \sigma_1\sigma_2\sigma_3 \,, \tag{7.15}$$

where σ_1, σ_2, σ_3 are the principal stresses.

Hydrostatic and Deviatoric Stresses. The hydrostatic stress (a stress invariant) is defined as

$$\sigma_\mathrm{h} = \frac{\sigma_{xx} + \sigma_{yy} + \sigma_{zz}}{3} \,. \tag{7.16}$$

The respective hydrostatic stress tensor is hence

$$\boldsymbol{\sigma}_{ij}^\mathrm{h} = \begin{pmatrix} \sigma_\mathrm{h} & 0 & 0 \\ 0 & \sigma_\mathrm{h} & 0 \\ 0 & 0 & \sigma_\mathrm{h} \end{pmatrix} = \sigma_\mathrm{h} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \,. \tag{7.17}$$

It is an experimental observation that the hydrostatic stress does not initiate any plastic deformation. However, the ultimate amount of plastic deformation that a metal can withstand increases with increasing compressive hydrostatic stress. Plastic deformation is induced by the deviatoric stress tensor obtained by

$$\boldsymbol{\sigma}_{ij}' = \boldsymbol{\sigma}_{ij} - \boldsymbol{\sigma}_{ij}^\mathrm{h} = \begin{pmatrix} (\sigma_{xx} - \sigma_\mathrm{h}) & \tau_{xy} & \tau_{xz} \\ \tau_{yx} & (\sigma_{yy} - \sigma_\mathrm{h}) & \tau_{yz} \\ \tau_{zx} & \tau_{zy} & (\sigma_{zz} - \sigma_\mathrm{h}) \end{pmatrix} \,. \tag{7.18}$$

Flow Condition

The flow condition (also called yield condition) is a hypothesis describing the condition that must be fulfilled for plastic flow to start or to pursue. Most of the flow conditions are in terms of the components of the stress tensor. Any flow condition must fulfil the trivial case of the uniaxial stress state for which plastic flow starts if the axial stress is equal to the yield stress. In metal forming, instead of the yield stress the term flow stress is preferred. Referring to Fig. 7.44 the elastic region is terminated by the initial yield stress or the initial flow stress $\sigma_{\mathrm{f}0}$. Increasing the tension force further will also increase the axial stress, however with a much lower gradient than in the elastic region. If loading is terminated at a flow stress of $\sigma_{\mathrm{f}1}$ and the specimen is unloaded and then reloaded, plastification will this time start at a higher flow stress of $\sigma_{\mathrm{f}1}$. This phenomenon is called strain or work hardening. Under certain circumstances the flow stress $\sigma_{\mathrm{f}1}$ may be equal to the initial flow stress $\sigma_{\mathrm{f}0}$. Such kinds of materials are called non-hardening or perfect-plastic materials. If the flow stress after plastification decreases the material is strain softening material.

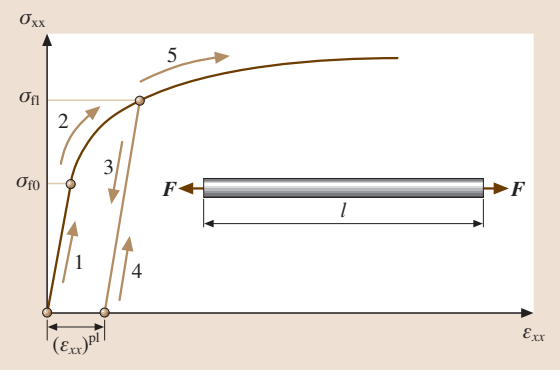


Fig. 7.44 Flow stress in simple tension test (numbers give the sequence)

The two flow conditions used extensively in metal forming for general three-dimensional stress cases are the Tresca flow condition and the von Mises flow condition.

The Tresca Flow Condition. The Tresca flow condition states that plastic flow will start or pursue if the maximum shear stress $\tau_{\max}$ is equal to a material constant that can be found as $\sigma_\mathrm{f}/2$ after tuning this criterion by the simple tension test

$$\tau_{\max} = \frac{\sigma_\mathrm{f}}{2} \,. \tag{7.19}$$

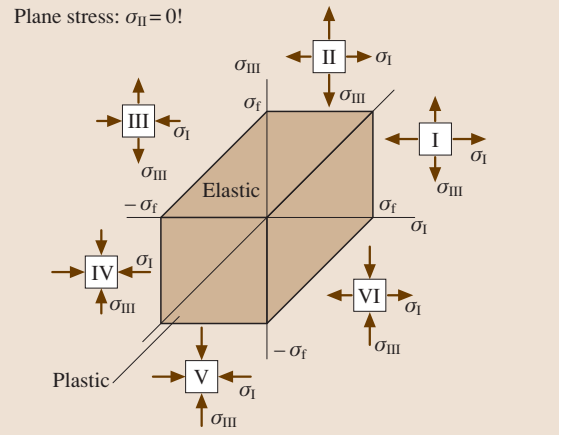
Or, in terms of principal normal stresses

$$\sigma_{\max} - \sigma_{\min} = \sigma_\mathrm{f} \,, \tag{7.20}$$

where $\sigma_{\max}$ and $\sigma_{\min}$ are the largest and smallest principal normal stresses. Figure 7.45 shows the flow locus for plane stress states according to Tresca. The table in the figure depicts the various forms that the Tresca flow criterion takes in terms of the spatial plane stress components σ_I and σ_III.

The von Mises Flow Condition. Different to the Tresca flow condition the von Mises flow condition considers all the principal shear stresses. Accordingly, plastic flow starts or pursues if

$$c' = \sqrt{\frac{1}{3}\left[\left(\tau_{\max}^1\right)^2 + \left(\tau_{\max}^2\right)^2 + \left(\tau_{\max}^3\right)^2\right]} \,, \tag{7.21}$$



Region		
I	$\sigma_I > \sigma_{III} > 0$	$\sigma_f = \sigma_I$
II	$\sigma_{III} > \sigma_I > 0$	$\sigma_f = \sigma_{III}$
III	$\sigma_{III} > 0 > \sigma_I$	$\sigma_f = \sigma_{III} - \sigma_I$
IV	$0 > \sigma_{III} > \sigma_I$	$\sigma_f = -\sigma_I$
V	$0 > \sigma_I > \sigma_{III}$	$\sigma_f = -\sigma_{III}$
VI	$\sigma_I > 0 > \sigma_{III}$	$\sigma_f = \sigma_I - \sigma_{III}$

Fig. 7.45 Tresca flow hexagon

where c' is a material constant. Again tuning this equation by the simple tension test supplies

$$\sigma_f = \sqrt{\frac{1}{2}\left[\begin{array}{l}(\sigma_{xx}-\sigma_{yy})^2+(\sigma_{yy}-\sigma_{zz})^2\\ \quad +(\sigma_{zz}-\sigma_{xx})^2+6\left(\tau_{xy}^2+\tau_{yz}^2+\tau_{zx}^2\right)\end{array}\right]}\,. \tag{7.22}$$

Figure 7.46 shows the flow locus for plane stress states. In this case, (7.22) reduces to

$$\sigma_f = \sqrt{\sigma_I^2 + \sigma_I\sigma_{III} + \sigma_{III}^2}\,. \tag{7.23}$$

A comparison of the von Mises flow criterion with the Tresca criterion reveals that for plane stress states the maximum deviation between the two criteria is 15.5%. Each criteria finds a different shear stress in pure shear for initiating plastic flow k

$$k = \begin{cases} \frac{1}{2}\sigma_f & \text{by Tresca} \\ \frac{1}{\sqrt{3}}\sigma_f & \text{by von Mises}\,. \end{cases} \tag{7.24}$$

Both criteria assume that plastic flow is independent of the hydrostatic stress and that plastic flow is independent of the sense of the stresses. Violation of the latter is also called the Bauschinger effect or kinematic hardening.

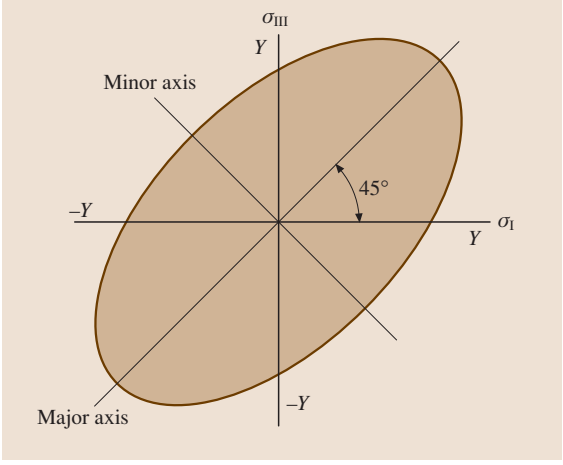


Fig. 7.46 Von Mises flow ellipse

Flow Rule

The relationship between plastic strains and stresses is expressed by the Levy–Mises flow rule. Unlike in Hooke's law for elastic deformations, in plasticity true strain rates are related to deviatoric stresses

$$\dot{\boldsymbol{\varepsilon}}_{ij} = \lambda\boldsymbol{\sigma}'_{ij}\,, \tag{7.25}$$

where λ is a nonnegative real number. By inserting (7.25) into (7.22) the nonnegative parameter can be eliminated, yielding

$$\dot{\boldsymbol{\varepsilon}}_{ij} = \frac{\sqrt{3I_2^{\dot{\varepsilon}_{ij}}}}{\sigma_f}\boldsymbol{\sigma}'_{ij}\,, \tag{7.26}$$

where $I_2^{\dot{\varepsilon}_{ij}}$ is the second invariant of the strain rate tensor.

Equivalent Plastic Strain and the Flow Curve

The plastic power per unit volume due to the stresses during plastic deformation can be expressed by

$$p = \sum_i^3\sum_j^3 \sigma_{ij}\dot{\varepsilon}_{ij} = \sum_i^3\sum_j^3 \sigma'_{ij}\dot{\varepsilon}_{ij}\,. \tag{7.27}$$

Equating this power to the power done by the respective flow stress and a fictitious strain rate leads to the definition of the equivalent plastic strain rate $\dot{\bar{\varepsilon}}$

$$p = \sum_i^3\sum_j^3 \sigma'_{ij}\dot{\varepsilon}_{ij} = \sigma_f\dot{\bar{\varepsilon}}\,. \tag{7.28}$$

Together with (7.26) this yields

$$\dot{\bar{\varepsilon}} = \sqrt{\frac{4}{3} I_2^{\dot{\varepsilon}_{ij}}} = \sqrt{\frac{2}{3}\left(\sum_i^3 \sum_j^3 \dot{\varepsilon}_{ij}\dot{\varepsilon}_{ij}\right)}$$
$$= \sqrt{\frac{2}{3}\left[\left(\dot{\varepsilon}_{xx}^2 + \dot{\varepsilon}_{yy}^2 + \dot{\varepsilon}_{zz}^2\right) + 2\left(\dot{\varepsilon}_{xy}^2 + \dot{\varepsilon}_{yz}^2 + \dot{\varepsilon}_{xz}^2\right)\right]} . \quad (7.29)$$

The total equivalent plastic strain is the time integral of (7.29)

$$\bar{\varepsilon} = \int_t \dot{\bar{\varepsilon}}\, \mathrm{d}t , \quad (7.30)$$

or in terms of total strain components

$$\bar{\varepsilon} = \sqrt{\frac{2}{3}\left(\dot{\varepsilon}_1^2 + \dot{\varepsilon}_2^2 + \dot{\varepsilon}_3^2\right)} . \quad (7.31)$$

The equivalent plastic strain is a measure of the increase of dislocation density and the mutual hindering of dislocations, i. e. it measures the increase of flow stress. The change of flow stress with the equivalent plastic strain is a material property called the flow curve

$$\sigma_\mathrm{f} = \sigma_\mathrm{f}(\bar{\varepsilon}) . \quad (7.32)$$

Typical flow curves for various temperatures and strain rates are shown in Fig. 7.47 for the typical cold forming steel C15. Usually, the flow stress at a given equivalent strain decreases with temperature and increases with the strain rate. The flow curve can be determined basically in simple tension, compression, and torsion tests. All theses tests have their limits regarding attainable strain values and stress states [7.29].

Temperature Increase During Forming

During plastic deformation about 85 to 90% of the plastic work is dissipated as heat. For a homogenous adiabatic deformation the average temperature increase can be computed by

$$\Delta T = \frac{(0.85\text{–}0.95)\bar{\sigma}_{\mathrm{f,mean}}\bar{\varepsilon}}{\rho c} , \quad (7.33)$$

where $\bar{\sigma}_{\mathrm{f,mean}}$ is the mean flow stress during deformation, $\bar{\varepsilon}$ is the mean equivalent plastic strain, c is the specific heat capacity, and ρ is the density of the material. Table 7.19 shows the average temperature increase for a moderate equivalent plastic strain of unity for various materials.

Analytical Methods

All methods discussed in this section assume rigid-plastic behavior and quasi static loading [7.30–32].

Elementary Methods of Plasticity and Friction Models.

Here only the equilibrium methods are discussed. The equilibrium methods are based on the static equilibrium of simple, elementary, standard free-bodies (semiinfinitesimal slabs) for which the strain-rate state can be easily determined. Basically three types of elementary free-bodies are used: rectangular slab (Fig. 7.48a), circular slab (Fig. 7.48b), and tubular slab (Fig. 7.48c). The basic assumptions are:

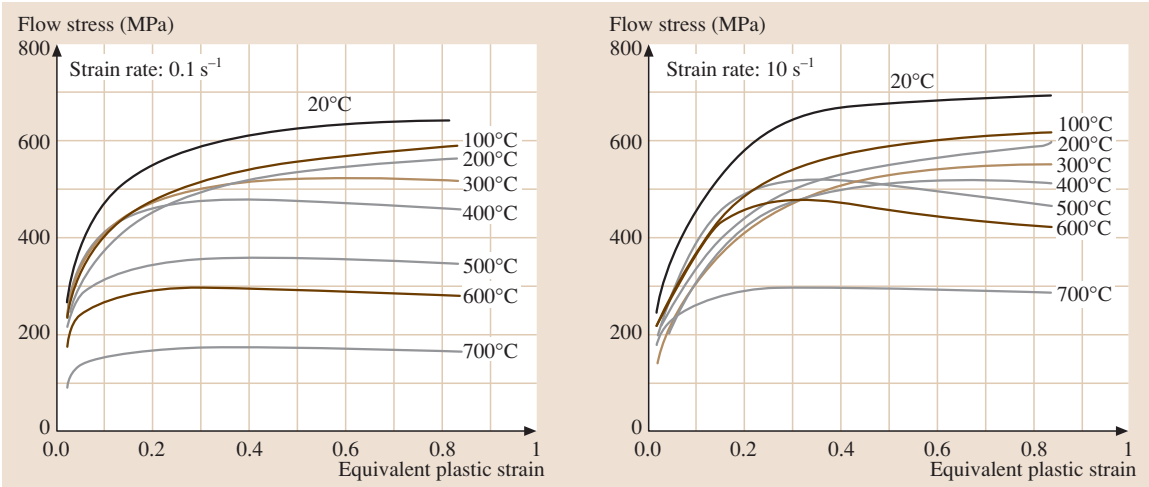


Fig. 7.47 Flow curve of a low carbon steel for various temperatures and strain rates for a C15 steel (after [7.28])

Table 7.19 Average temperature increases during a homogeneous metal forming process with average plastic strain of unity (after [7.33])

	ρ (kg/dm^3)	c (J/(kg K))	$\sigma_{f,mean}$ (N/mm^2)	ΔT (°C)
Steel	≈ 7.8	≈ 500	400–1200	100–300
Al-alloys	≈ 2.7	≈ 1000	100–300	35–100
Cu-alloys	≈ 8.7	≈ 400	200–400	50–100
Ti-alloy	≈ 4.5	≈ 600	750–1500	250–500

1. The strain-rate state in each semiinfinitesimal slab is homogenous.
2. Plane sections remain plane.
3. The stresses along the characteristic directions of the slabs are principal stresses.
4. Friction at the contact interfaces is described either by the Coulomb friction model or by the constant shear model.

The Coulomb friction model states that the frictional shear stress at a surface element at the interface is proportional to the normal stress

$$\tau_{\text{friction}} = \mu \sigma_{\text{normal}} \,. \tag{7.34}$$

The proportionality factor μ is called the coefficient of friction. In metal forming the range for μ is given as

$$0 \le \mu \le 0.577 \,. \tag{7.35}$$

The constant shear friction model states that the frictional stress is proportional to the flow stress in shear k

$$\tau_{\text{friction}} = mk \,. \tag{7.36}$$

The proportionality factor m is called the friction factor. The range of m is given by

$$0 \le m \le 1 \,. \tag{7.37}$$

Both models only approximately describe reality. Recent friction models are a combination of both elementary models. Typical values for the coefficient of friction and the friction factor are given in Table 7.20.

The unknown stresses can be solved by the equilibrium equations and the flow condition. The flow rule is not used. The accuracy of this approach is less, the larger the friction or the inclination of the free surfaces are.

Upper and Lower Bound Methods. The upper bound method is based on the first law of thermodynamics. For a kinematically admissible velocity field that satisfies volume constancy and all velocity boundary conditions of a plastically deforming body, an upper bound for the forming force is given by [7.34]

$$F_{\text{tool}}\big|_{\text{upper bound}} = \frac{1}{v_{\text{tool}}} \left(\int_V \sigma_f \dot{\bar{\varepsilon}} \, \mathrm{d}V + \int_{A_s} mk v_{\text{shear}} \, \mathrm{d}A - \int_{A_f} \boldsymbol{p}\boldsymbol{v} \, \mathrm{d}A \right), \tag{7.38}$$

where v_{tool} is the tool velocity, $\boldsymbol{p}$ is the known pressure vector acting on the workpiece, v_{shear} is the shearing velocity at shear surfaces, $\boldsymbol{v}$ is the velocity vector of points on the workpiece surface, A_s is the workpiece area on

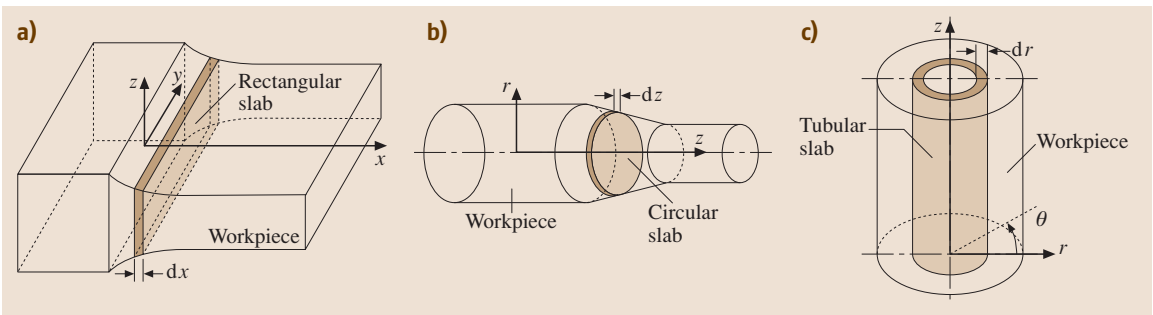


Fig. 7.48a–c Various slab elements: **(a)** rectangular slab, **(b)** circular slab, **(c)** tubular slab

Table 7.20 Typical values for coefficient of friction and friction factor

Process	Coefficient of friction μ	Friction factor m
Cold forging (steel, stainless steel, Cu-alloys, brass)	0.05–0.10 (lower values for phosphated workpieces)	0.05–0.10
Cold forging (Al-, Mg-alloys)	0.05	0.05–0.10
Wire drawing (steel, stainless steel, Cu-alloys, brass)	0.05–0.10	–
Wire drawing (Al-, Mg-alloys)	0.03–0.10	–
Hot forging	Use not recommended	0.20–0.40
Forging of Ti and Ni alloys	Use not recommended	0.10–0.30 (glass lubrication)
Hot rolling	Use not recommended	0.70–1.00 (no lubrication)
Deep drawing (steel)	0.05–0.10	Use not recommended
Deep drawing (stainless steel)	0.10	Use not recommended
Deep drawing (Cu-alloys, brass)	0.05–0.10	Use not recommended
Deep drawing (Al-, Mg-alloys)	0.05	Use not recommended
Ironing (steel)	0.05–0.10	Use not recommended
Ironing (stainless steel)	0.05	Use not recommended
Ironing (Cu-alloys, brass)	0.10	Use not recommended
Ironing (Al- and Mg-alloys)	0.05	Use not recommended

which friction is specified, and A_f is the workpiece area on which forces are specified. To apply the upper bound method the plastic region has to be estimated, the constant flow stress has to be estimated in the plastic region, friction stresses must be constant and hence described by the constant shear model, and finally, a kinematically admissible velocity field or appropriate shear planes have to be assumed.

Lower bound methods lead to forces that are always lower than the actual forces. Here, a statically admissible stress field has to be guessed that fulfils the equilibrium equations and the stress boundary conditions. Ideally, for an analysis the force can be limited between an upper and lower bound. Yet, the application of the lower bound method is much more difficult since the guess of an admissible stress field is not straightforward. Furthermore, in practical applications an upper bound for the forming forces are sufficient. The *upper bound* property for the forming force is given if and only if the material's flow stress is correct and the friction stresses are correct. Otherwise, the computed *upper bound* may also be *lower* than the actual physical forces.

The Slip Line Field Solution. The slip line field solution also assumes rigid perfectly plastic material behavior. Furthermore, the plane strain state is assumed. Moreover, the processes have to be frictionless. If these assumptions are fulfilled, then the theory supplies the exact solution. The slip line field solution is based on the governing equations including the flow condition, the volume constancy equation, the flow rule, and the equilibrium equations for the plane strain state. These equations build up a hyperbolic system of partial differential equations that can be solved by the method of characteristics. If the stresses are expressed in terms of the hydrostatic stress and the orientation of the stress element in the maximum shear stress direction, the two characteristics lines are orthogonal to each other (α and β-slip lines) and correspond to the directions of maximum shear stress; therefore they are called slip lines. The governing equations can be written as ordinary differential equations along each slip line. These equations can be summarized as the Hencky equations for the stresses

$$\begin{aligned} \sigma_h - 2k\phi &= \text{constant along the } \alpha\text{-slip line,} \\ \sigma_h + 2k\phi &= \text{constant along the } \beta\text{-slip line,} \end{aligned} \qquad (7.39)$$

and the Geiringer equations in terms of the particle velocities

$$\begin{aligned} \frac{dv_\alpha}{ds} &= v_\beta \frac{d\phi}{ds} \quad \text{along } \alpha\text{-slip line,} \\ \frac{dv_\beta}{ds} &= v_\alpha \frac{d\phi}{ds} \quad \text{along } \beta\text{-slip line,} \end{aligned} \qquad (7.40)$$

where ϕ is the inclination of the slip line and s is distance along the slip line. Constructing the slip line field from known boundary conditions, using (7.39) and

(7.40) supplies the stress and velocity (and hence strain rates) at any point in the deformation zone.

Empirical Methods

The Visio-Plasticity Method. The visio-plasticity consists of splitting the workpiece before deformation into two parts and marking a grid on one side, then closing these two parts and forming the workpiece. After forming the grids are deformed and are measured out in order to determine the particle velocities and their displacements. From this data the strain rates and then the stresses can be computed. Figure 7.66 shows a typical visio-plasticity specimen used in forward rod extrusion. The basic assumption is that the initial plane on which the grid produced remains plane during plastic deformation. The grids can be marked on the specimens by mechanical methods, photomechanical methods, electrochemical methods, and photolithographical methods.

The velocities can be determined in the case of steady state processes (Fig. 7.49) by first computing a time increment

$$\Delta t = \frac{u_{z0}}{\dot{u}_{z0}} = \frac{u_z}{\dot{u}_z(\bar{r}_B, \bar{z}_B)} = \frac{u_r}{\dot{u}_r(\bar{r}_B, \bar{z}_B)} , \qquad (7.41)$$

where $\bar{r}_B$ and $\bar{z}_B$ are the mean coordinates of point B. Hence the velocity components of the point B can be found as

$$\dot{u}_z(\bar{r}_B, \bar{z}_B) = \frac{u_z}{\Delta t} = \frac{u_z}{u_{z0}}\dot{u}_{z0} ,$$
$$\dot{u}_r(\bar{r}_B, \bar{z}_B) = \frac{u_r}{\Delta t} = \frac{u_r}{u_{z0}}\dot{u}_{z0} . \qquad (7.42)$$

Using these components the strain rate components can be found. The deviatoric stress components are given by the flow rule (7.26). The hydrostatic stress found from stress boundary conditions and the previously found deviatoric stresses are used to determine the total stress components at the end.

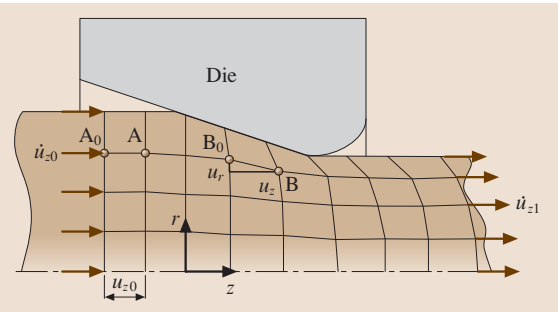


Fig. 7.49 Visio-plasticity analysis (after [7.33])

Physical Simulation. Using modeling materials like wax or plastiline the flow pattern can be visualized and observed through transparent dies for instance [7.35]. The key idea here is the fact that two materials behave similarly from the viewpoint of plasticity if they have the same strain hardening exponent n. It is possible to add to wax and plastiline additives such as Kaolin, Indramic, and Filia to change the n value and even the strain-rate sensitivity value m. Also, the ductility and the strength level can be manipulated by additives. Hence, the basic difference to the forming process of metals is the much lower stress level.

Numerical Methods

The governing equations of plasticity cannot be solved in closed form except for very few simple applications. The approximate solution of the differential equations can be performed numerically by various methods. Among these methods the finite element solution has found the most extended application because of its generality and efficiency.

Rigid-Plastic Finite Element Formulations. The elastic part of the strain is neglected and the whole body is assumed to deform rigid-plastically. The solutions obtained by these formulation are not able to supply residual stresses, springback, and similar elastic strain dependent variables. Furthermore, since the formulation assumes a pseudo-elastic behavior in the rigid parts of the workpiece, friction in these parts may be modeled wrongly [7.36].

The rigid-plastic formulation is based on the Markov's variational principle that holds only for the rigid-plastically deforming bodies

$$\Pi(v) = \int_V \sigma_f \dot{\bar{\varepsilon}}\, dV + \int_V \frac{S}{2}\dot{\varepsilon}_V^2\, dV - \int_{A_t} \boldsymbol{t}\boldsymbol{v}\, dA$$
$$\Rightarrow \text{stationary}, \qquad (7.43)$$

where Π is the functional, $\boldsymbol{v}$ the velocity vector of material particles, $\dot{\varepsilon}_V$ the volumetric strain rate, $\boldsymbol{t}$ the specified traction acting over the surface A_t of the body, and S a constant number called the penalty factor. The penalty factor term is necessary to impose the volume constancy constraint approximately. Alternatively, the volume constancy constraint can be imposed by the Lagrange multiplier λ exactly.

It can be shown that the hydrostatic stress is given in both cases by

$$\sigma_H = S\dot{\varepsilon}_V ,$$
$$\sigma_H = \lambda . \qquad (7.44)$$

These functionals can be easily discretized by assuming shape functions for the velocities over the element domain.

After the standard discretization procedure the resulting finite element equations read

$$\left(\mathbf{K}_{\mathrm{D}}^{e}+\mathbf{K}_{\mathrm{H}}^{e}\right)\boldsymbol{v}^{e}=\boldsymbol{f}^{e}\,. \tag{7.45}$$

$\boldsymbol{f}^{e}$ is the nodal force vector of the element compatible with the nodal velocity vector $\boldsymbol{v}^{e}$. The nonlinear deviatoric stiffness matrix is given by $\mathbf{K}_{\mathrm{D}}^{e}$ and the linear hydrostatic stiffness matrix by $\mathbf{K}_{\mathrm{H}}^{e}$. The resulting nonlinear (w.r.t. nodal velocities) equations (7.45) can be solved by standard numerical methods. The common ones applied are the direct iteration and the Newton–Raphson method. Both methods are iterative and are applied in increments. The time integration is performed explicitly in most commercial software.

Implicit Static Elasto-Plastic Finite Element Formulations. These formulations usually assume an additive composition of the elastic and plastic strain rate tensors

$$\dot{\boldsymbol{\varepsilon}}_{ij}\approx\dot{\boldsymbol{\varepsilon}}_{ij}^{\mathrm{el}}+\dot{\boldsymbol{\varepsilon}}_{ij}^{\mathrm{pl}}\,. \tag{7.46}$$

This is based on so-called hypoelastic models. Also models based on hyperelasticity are used that lead to a multiplicative split of elastic and plastic deformations. For the elastic strain rates the generalized Hooke's law and for the plastic strain rates the Levy–Mises equations are used. These then lead to the modified Prandtl–Reuss equations between the objective (frame-invariant) rate of the stress tensor and the strain rates

$$\hat{\boldsymbol{\sigma}}_{ij}=\mathbf{C}_{ijmn}\dot{\boldsymbol{\varepsilon}}_{mn}\,. \tag{7.47}$$

Various types of objective rates can be used such as the Jaumann rate, the Truesdell rate, or generally any Lie-derivative of the true stress tensor. The constitutive fourth order tensor contains the elastic constants and the plastic properties such as the normal of the flow surface and the slope of the flow curve. For consistent linearization (7.47) has to be modified slightly.

The elasto-plastic field equations are derived from the principle of virtual work supplying (neglecting inertial forces)

$$\int_{V}\boldsymbol{\sigma}_{ij}\delta\left(\frac{\partial u_j}{\partial x_i}\right)\mathrm{d}V-\int_{A}t_i\delta u_i\,\mathrm{d}A=0\,. \tag{7.48}$$

Equation (7.48) must be fulfilled at the unknown current configuration. Linearization of this equation about the last known state and space discretization supply the nonlinear finite element equations. Time integration is performed primarily implicitly.

Explicit Dynamic Elasto-Plastic Finite Element Formulations. The explicit dynamic finite element formulations are based on the virtual work principle to which an inertia term is added

$$\int_{V}\boldsymbol{\sigma}_{ij}\delta\left(\frac{\partial u_j}{\partial x_i}\right)\mathrm{d}V-\int_{A}t_i\delta u_i\,\mathrm{d}A+\int_{V}\rho\ddot{\boldsymbol{u}}_i\delta u_i\,\mathrm{d}V=0\,, \tag{7.49}$$

where $\ddot{\boldsymbol{u}}_i$ is the acceleration vector and ρ the density. Discretization of (7.49) leads to

$$\mathbf{M}\ddot{\boldsymbol{u}}=\boldsymbol{F}-\boldsymbol{I}\,. \tag{7.50}$$

Here, $\mathbf{M}$ is the mass matrix, $\boldsymbol{F}$ the external force vector, and $\boldsymbol{I}$ the internal force vector at the current time. Time discretization by a central difference scheme and by adding a damping term $\mathbf{C}$ supplies the fundamental equations for the formulation

$$\left[\frac{1}{(\Delta t)^2}\mathbf{M}\right]\boldsymbol{u}^{t+\Delta t}=\boldsymbol{F}^{t}-\boldsymbol{I}^{t}+\left(\frac{\mathbf{M}}{2\Delta t}-\mathbf{C}\right)\left(\frac{\boldsymbol{u}^{t}-\boldsymbol{u}^{t-\Delta t}}{\Delta t}\right)\,. \tag{7.51}$$

Here Δt is the time increment for the computation and must satisfy the stability condition

$$\Delta t\le\frac{L}{c_{\mathrm{d}}}\,, \tag{7.52}$$

where c_{d} is the current dilatational wave speed of the material (speed of sound in that material) and L is the characteristic element dimension, which can be taken as the minimum distance between any two nodes of an element. The elastic wave speed can be found from

$$c_{\mathrm{d}}=\sqrt{\frac{2G(1-\nu)}{(1-2\nu)\rho}}\,. \tag{7.53}$$

7.2.4 Bulk Forming Processes

This section describes the basic bulk forming processes.

Upsetting

A workpiece of initial diameter d_0 and initial height h_0 is reduced between two flat dies to a specimen with final diameter d_1 and final height h_1 (Fig. 7.50). This process is also called open die forging or free forming.

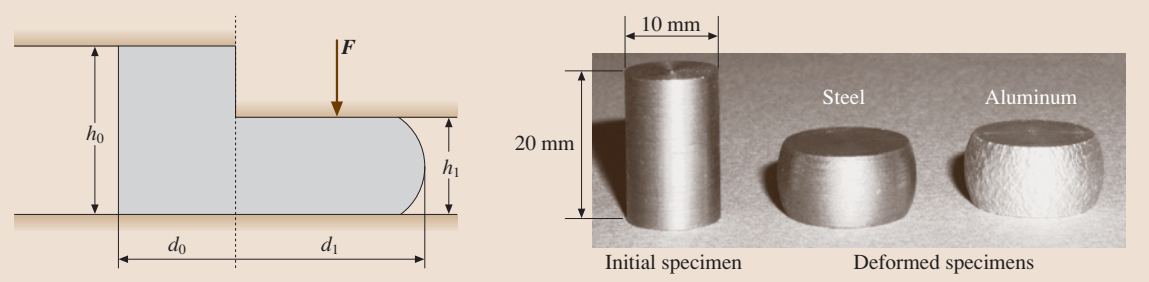


Fig. 7.50 **(a)** Principle of upsetting process. **(b)** Application examples

Due to friction between the dies and the workpiece the workpiece bulges. The upsetting process is used as a material characterization process and as a production process in various forms. The basic parameters characterizing the process are the ideal equivalent plastic strain

$$\bar{\varepsilon} = \ln\left(\frac{h_0}{h_1}\right) , \tag{7.54}$$

the upsetting ratio

$$s = \frac{h_0}{d_0} , \tag{7.55}$$

the relative height reduction

$$e = \frac{\Delta h}{h_0} = \frac{h_0 - h_1}{h_0} , \tag{7.56}$$

and finally ideal equivalent plastic strain rate

$$\dot{\bar{\varepsilon}} = \frac{\mathrm{d}\bar{\varepsilon}}{\mathrm{d}t} = \frac{v_{\mathrm{tool}}}{h} , \tag{7.57}$$

where v_{tool} is the tool velocity (precisely the velocity difference of the upper and lower dies).

Utilizing the elementary plasticity theory (slab method), the axial, radial and tangential stresses are found for Coulomb friction as

$$\sigma_z(r) = -\sigma_{\mathrm{f}} \mathrm{e}^{\frac{2\mu}{h}\left(\frac{D}{2}-r\right)} \approx \underbrace{-\sigma_{\mathrm{f}}\left[1+\frac{2\mu}{h}\left(\frac{D}{2}-r\right)\right]}_{\text{after a Taylor's series expansion}} ,$$

$$\sigma_r(r) = \sigma_\theta(r) = \sigma_z(r) + \sigma_{\mathrm{f}} \approx \underbrace{-\sigma_{\mathrm{f}}\left[\frac{2\mu}{h}\left(\frac{D}{2}-r\right)\right]}_{\text{after a Taylor's series expansion}} , \tag{7.58}$$

and for the constant shear friction model

$$\sigma_z(r) = -\sigma_{\mathrm{f}}\left[1+\frac{m}{\sqrt{3}}\frac{D}{h}\left(1-\frac{r}{D/2}\right)\right] ,$$

$$\sigma_r(r) = \sigma_\theta(r) = -\sigma_{\mathrm{f}}\frac{m}{\sqrt{3}}\frac{D}{h}\left(1-\frac{r}{D/2}\right) . \tag{7.59}$$

The stress distribution is shown in Fig. 7.51 for the linearized Coulomb friction model. The upsetting force at a generic height is given by

$$|F_z| = -\int_A \sigma_z \,\mathrm{d}A \approx \frac{\pi}{4} D^2 \sigma_{\mathrm{f}} \left(1+\frac{1}{3}\mu\frac{D}{h}\right) . \tag{7.60}$$

Figure 7.52 exhibits a typical force–displacement curve for upsetting. The steep increase of the force to-

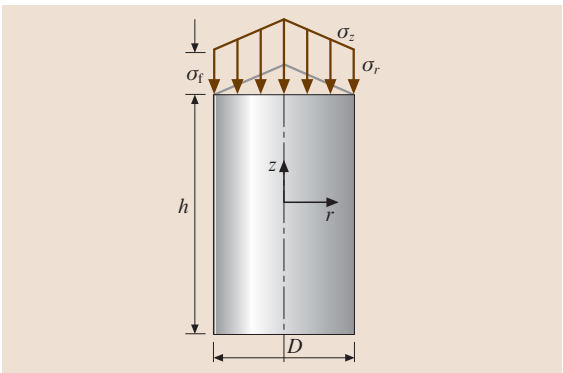


Fig. 7.51 Stresses in upsetting

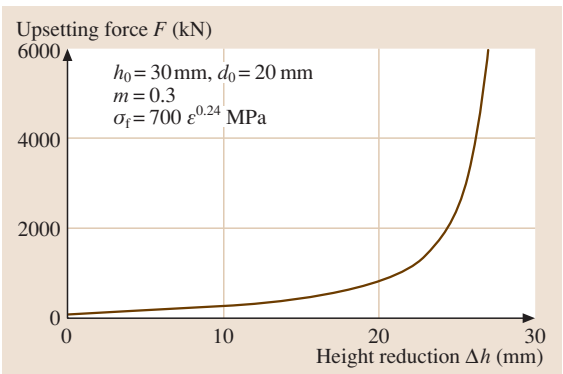


Fig. 7.52 Force displacement curve in upsetting

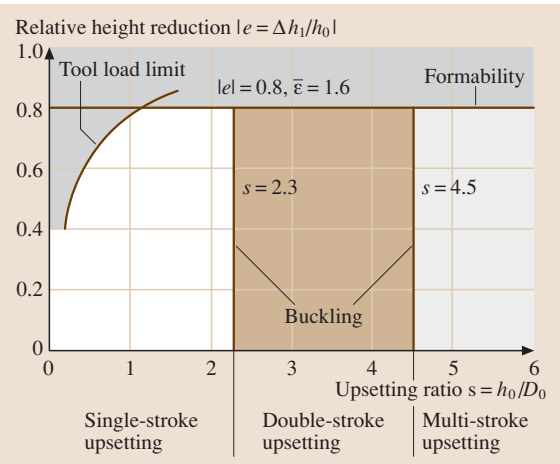


Fig. 7.53 Process limits of cold upsetting (after [7.22, 37])

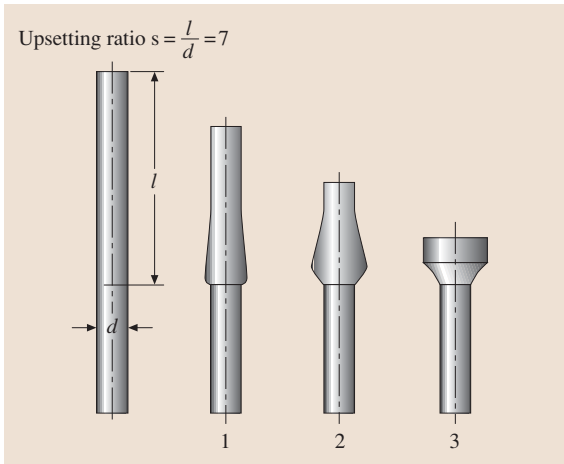


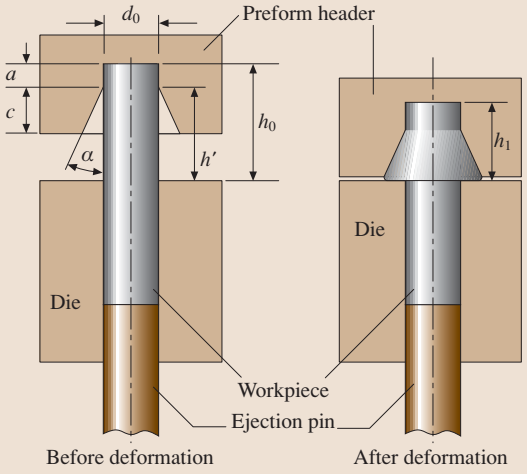
Fig. 7.54 Heading in 3 steps (after [7.37])

Table 7.21 Typical accuracy during mass production by upsetting (after [7.38])

Nominal dimension (mm)	5	10	20	30	40	50	100
Head height accuracy (mm)	0.18	0.22	0.28	0.33	0.38	0.42	0.50
Head diameter accuracy (mm)	0.12	0.15	0.18	0.20	0.22	0.25	0.30

Table 7.22 Some typical tool materials for cold upsetting and heading (after [7.37, 38])

Tool part	Material description by DIN	Material number	Hardness (HRC)
Dies	S6-5-2	1.3343	59–63
	X165CrMoV12	1.2601	60–62
	100V1	1.2833	58–61
	55NiCr10	1.2718	54–58
Shrink rings	56NiCrMoV7	1.2714	40–50
	X40CrMoV51	1.2344	40–50
	X2NiCoMo1885	1.6359	50–53
Preform punches (bulk)	100V1	1.2833	57–60
	145V33	1.2838	57–60
Preform punches (shrinked)	S6-5-2	1.3343	60–63
	X165CrMoV12	1.2601	60–63
End-form punches (bulk)	100V1	1.2833	58–61
	145V33	1.2838	58–61
End-form punches (shrinked)	S6-5-2	1.3343	60–63
	X165CrMoV12	1.2601	60–63
Ejector	X40CrMoV51	1.2344	55–58
	60WCrV7	1.2550	55–58



Pre-upsetting design

Upsetting ratio $s=h_0/d_0$	Cone angle 2α (°)	Guide length a (mm)	Conical portion c (mm)
2.5	15	0.6 d_0	1.37 d_0
3.3	15	1.0 d_0	1.56 d_0
3.9	15	1.4 d_0	1.66 d_0
4.3	20	1.7 d_0	1.56 d_0
4.5	25	1.9 d_0	1.45 d_0

Fig. 7.55 Preform design (after [7.39])

wards the smaller specimen heights is characteristic for upsetting.

Cold upsetting has the process limits shown in Fig. 7.53. These limits consist of the formability of the workpiece material characterized by a maximum of 80% height reduction, the buckling limit given by the maximum upsetting ratio of 2.3, and finally the tool load limit. The buckling limit can be increased by upsetting in multiple strokes. So, for a two-stroke process the upsetting ratio can be increased to 4.5, for instance. Formability limits are characterized by two types of cracks: shear cracks in the upsetted part and longitudinal cracks (basically due to initial faults in the workpiece).

For heading that is the upsetting of one end of the workpiece instead of the whole, Fig. 7.54 shows a three-step process necessitated by the buckling forming limit. Notice that the last step is no longer an open die process.

The design rules of the preforms during the head upsetting are given in Fig. 7.55.

Accuracy during cold upsetting depends on the conditions of the forming tools and machines. Wear of the tools is hereby an important issue. Table 7.21 gives some typical accuracy values for a reasonable number of parts formed.

A typical tool construction is given in Fig. 7.56. The die is usually supported by a shrink ring to increase its load carrying capacity. The basic tool materials are listed in Table 7.22 together with their hardness values.

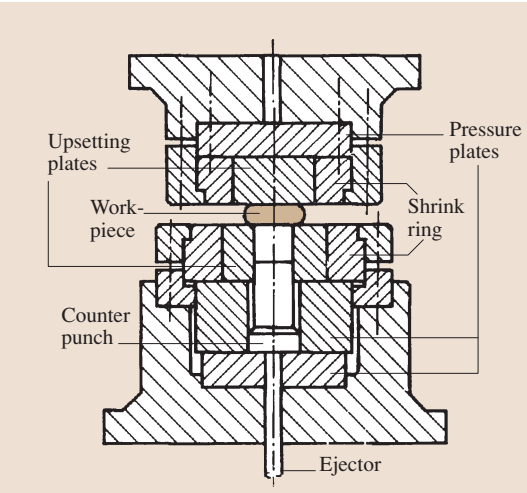


Fig. 7.56 Sketch of a typical upsetting tool (after [7.37])

Forging

Forging processes can be classified basically into two groups: Open-die forging and closed-die forging (Fig. 7.57). In open-die forging the dies possess simple geometries (usually flat) and the process is conducted hot. This process primarily serves to produce preforms for subsequent forming processes by reducing or increasing the cross-section of the workpiece. Upsetting is a fundamental process of open-die forging.

The basic closed-die forging process is impression forging in which a flash is produced that has to be separated from the workpiece after forging, [7.40]. The processes of fullering and gathering are used as pre-

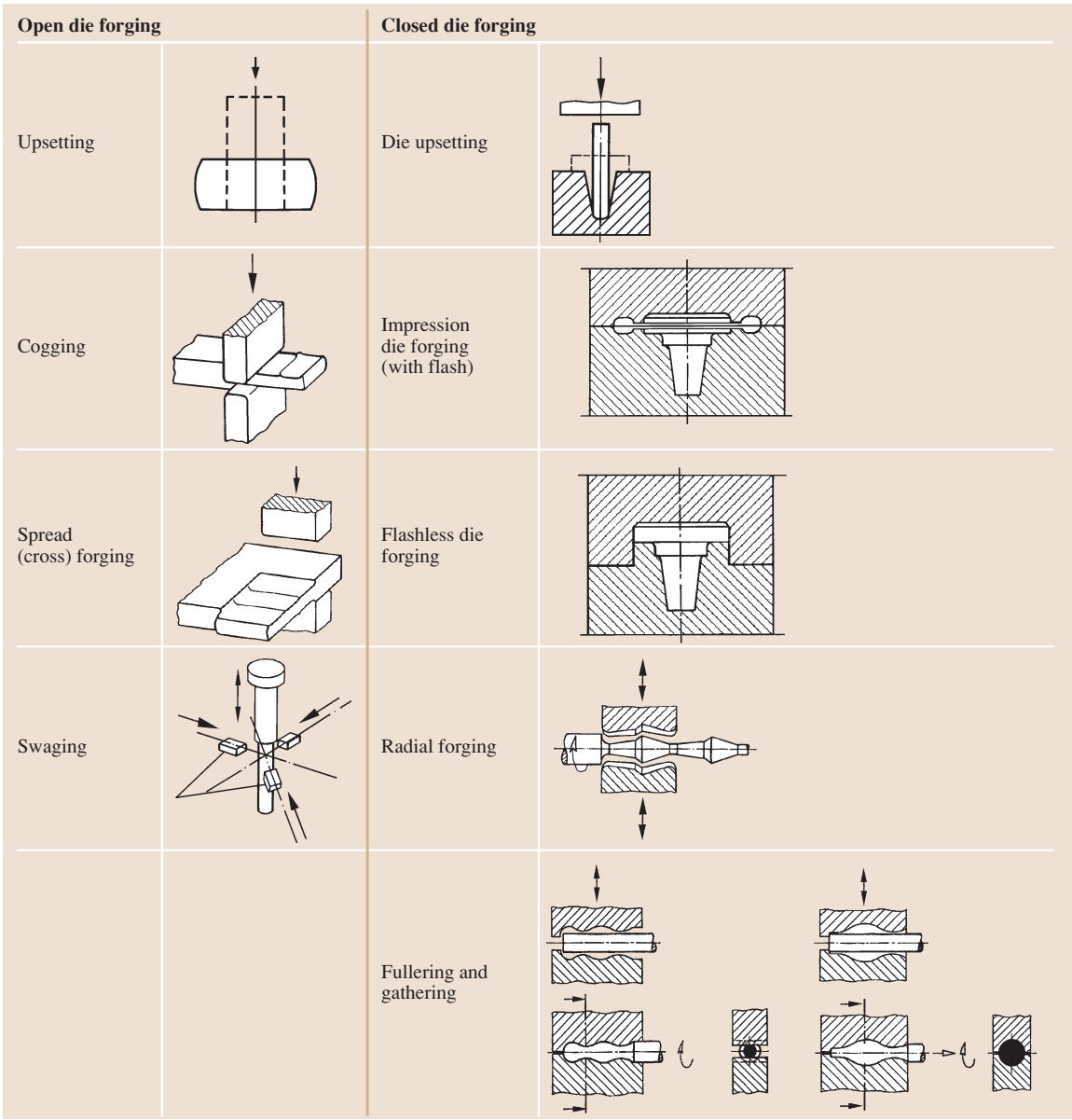


Fig. 7.57 Classification of forging processes

form processes before impression die forging. A typical impression forged product is shown in Fig. 7.58.

A typical closed die for forging (Fig. 7.59) consists of the upper and lower die parts. These parts form the forging cavity and the flash cavity separated by the parting line. A gutter is placed at the end of the flash land to collect the excess material. The forging cavity consists of webs and ribs. The flash is an important element in forging and controls the filling of the die. Furthermore, it serves as a means to compensate excessive material. On the other hand, it is the source of waste material in forging and should be kept as small as possible.

The force–displacement characteristics of closed-die forging exhibit three regions (Fig. 7.60): Until point P_1 the workpiece in the cavity is upset and then widens.

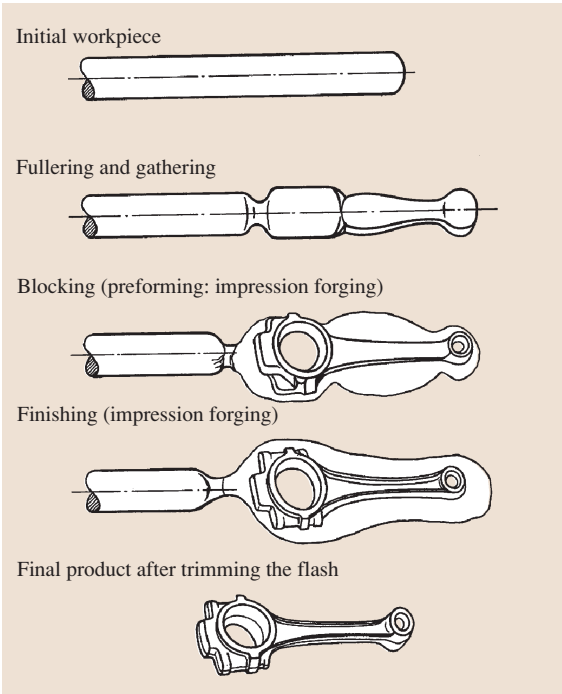


Fig. 7.58 Typical forging sequence (after [7.41])

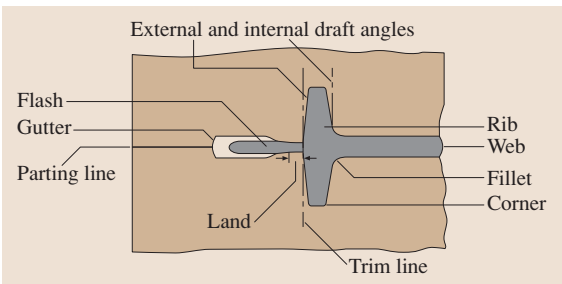


Fig. 7.59 Typical closed forging die (after [7.41])

At point P_1 the flash starts to form and hence the resistance to material flow increases, leading to the rising of the material during which the form of the die cavity is filled. At P_2 the die cavity fills and additional load is applied to close the dies finally. The necessary forming work is the shaded area under the load-displacement curve. This work has to be supplied as the forming energy by the forming machine.

The effect of the flash dimensions is twofold (Fig. 7.61): Increasing the flash–land ratio length/thickness (b/s), the braking effect of the flash increases and hence increasing the stress $\sigma_{z\,max}$ in the contact zone die-workpiece so that the forging force $\boldsymbol{F}$ increases. At

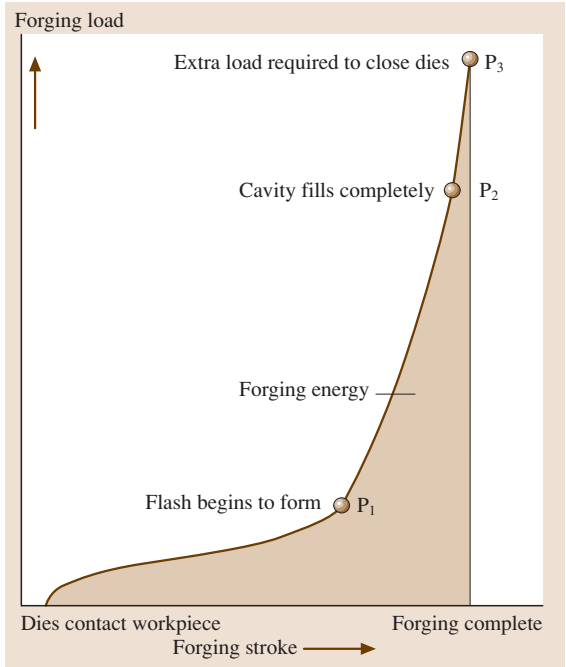


Fig. 7.60 Force–displacement characteristics of impression forging (after [7.42])

the same time however the excess mass Δm flowing into the die or the gutter decreases lowering the required forming energy W. In the given example, for a flash–land ratio of 4.5 there is nearly no change in the excess material so that the minimum forging energy is given for a die–land ratio of 4 to 5.

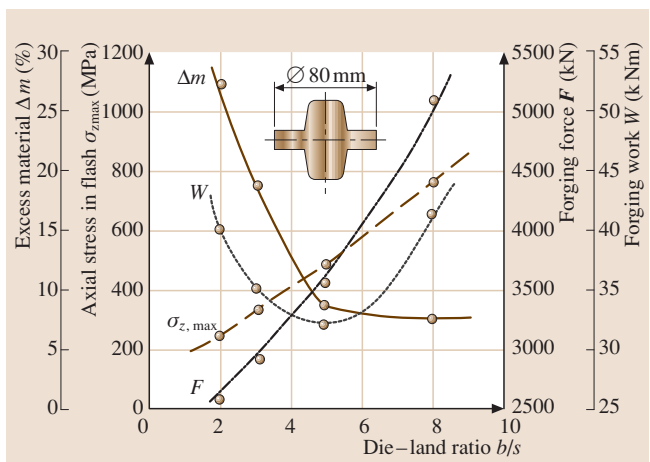


Fig. 7.61 Effect of flash dimensions on forging force and excessive material (after [7.37])

The stress distribution for axial-symmetric forgings resembles by elementary plasticity theory the stress distribution in upsetting (Fig. 7.62). The maximum stress occurs at the center of the part and is higher for wider flash–land ratios.

The maximum forging force can be estimated, for instance, by the upsetting equations as

$$\boldsymbol{F}^{\max} = -\sigma_{\mathrm{f}} \left(1 + 2\mu \frac{b}{s}\right) A_{\mathrm{p}} \,, \tag{7.61}$$

where σ_{f} is the flow stress (Table 7.23, see also [7.43]), μ the Coulomb friction coefficient (for steel at 1000 to 1100 °C $\mu = 0.12$ for lubricated and $\mu = 0.35$ for unlubricated cases), b/s (Table 7.24) the flash–land ratio, and A_{p} the total projected area including the flash. The flow stress can be approximated for the temperature and strain rate at the beginning of the process although the maximum force occurs towards the end. Hence, the strain rate is found as

$$\dot{\varepsilon} = \frac{v}{h_0} \,, \tag{7.62}$$

where h_0 is the initial height of the billet and v is the striking speed of the tool (Sect. 7.2.6). Notice that in Table 7.24 the b/s ratio is given for the three basic types of material flow in forging: upsetting, spreading (with large relative motions along the die surface), and rising (filling of the die cavities).

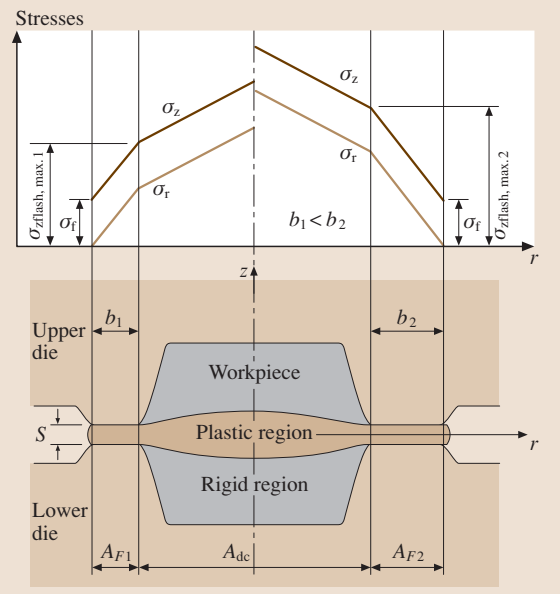


Fig. 7.62 Stress distribution in forging (after [7.37])

Table 7.23 Flow stress data for various hot forming metals, (after [7.38])

Material	m	K (MPa)	T (°C)
C15	0.154	99/84	1100/1200
C35	0.144	89/72	1100/1200
C45	0.163	90/70	1100/1200
C60	0.167	85/68	1100/1200
X10Cr13	0.091	105/88	1100/1250
X5CrNi189	0.094	137/116	1100/1250
X5CrNiTi189	0.176	100/74	1100/1250
E-Cu	0.127	56	800
CuZn28	0.212	51	800
CuZn37	0.201	44	750
CuZn40Pb2	0.218	35	650
CuZn20Al	0.180	70	800
CuZn28Sn	0.162	68	800
CuAl5	0.163	102	800
Al99.5	0.159	24	450
AlMn	0.135	36	480
AlCuMg1	0.122	72	450
AlCuMg2	0.131	77	450
AlMgSi1	0.108	48	450
AlMgMn	0.194	70	480
AlMg3	0.091	80	450
AlMg5	0.110	102	450
AlZnMgCu1.5	0.134	81	450

$\sigma_{\mathrm{f}}(T, \dot{\varepsilon}) = K(T)\dot{\varepsilon}^{m(T)}$

Typical failures in forged products are as follows:

1. Surface defects: These are surface scratches and cracks that are basically due to defects on the sheared surfaces of the billets or flash formation during preforming as well as imprints of cracks on the die surfaces.
2. Incomplete filling: Due to insufficient initial workpiece mass, inappropriate preforms, inaccurate workpiece positioning, wrong flash–land ratios, or too high gas pressures of the lubricant the die cavities may not fill completely.
3. Microstructural failures: The basic failure type is an insufficient grain flow.

Some of these failures can be removed by the subsequent machining operations of the forged products, so that these workpieces are not scrapped.

Basic die failures in forging are shown in Fig. 7.63. Dies are exposed to high temperatures, pressures, and sliding of workpiece materials along their sur-

Table 7.24 Recommended flash–land ratios (after [7.38])

Projected area of workpiece excluding the flash area (mm^2)	Flash–land ratio b/s		
	Mainly upsetting	Mainly spreading	Mainly rising
up to 2000	8	10	13
2001–5000	7	8	10
5001–10 000	5.5	6	7
10 001–25 000	4	4.5	5.5
26 000–70 000	3	3.5	4.5
71 000–150 000	2	2.5	3.5

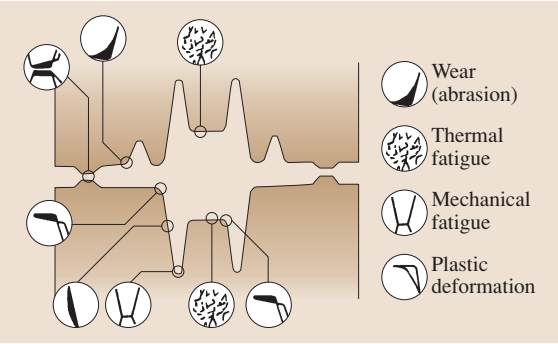


Fig. 7.63 Basic die failures in forging (after [7.37])

faces. The local pressures are as high as 1000 MPa or even higher and they vary with time. The wear is caused by sliding of workpiece materials with up to 50 m/s relative speed. Temperatures increase from 100–200 °C to 700–800 °C in extremely short durations with temperatures gradients of 1000 to 3000 °C/s. These effects lead to failure of dies through wear, thermal fatigue, mechanical fatigue, and plastic deformation.

Extrusion

Extrusion processes are classified according the technological differences into extrusion, impact extrusion and reducing (open-die extrusion) processes. The basic differences between extrusion and impact extrusion are that extrusion aims to produce profiles (semifinished products) whereas impact extrusion aims to produce single parts which are finished products. Furthermore, extrusion is performed usually hot (especially for steel materials almost always), whereas impact extrusion is usually performed cold (therefore also named as cold forging). Reducing is extrusion with area reductions less than 25 to 30%. In this case, there is no deformation of the material in the container. Further classifications are done according to the workpiece geometry and the relative motion between tools and workpiece during forming.

Figure 7.64 gives an overview of the principles of various extrusion processes.

In Fig. 7.65 the stress states as estimated by the elementary theory of plasticity for forward rod extrusion are given. Plastic deformation only occurs in the die-shoulder region and all stress components are compressive. The workpiece within the container remains basically elastic as well as the extruded rod after the exit of the die.

The fundamental parameters of the process are defined as area reduction

$$e = \frac{A_0 - A}{A_0} , \tag{7.63}$$

where A_0 and A_1 are the initial and final cross-sectional areas of the workpiece, and the ideal equivalent plastic strain φ

$$\varphi = \ln \frac{A_0}{A} = 2 \ln \frac{d_0}{d_1} . \tag{7.64}$$

The latter is the plastic equivalent strain at the axis of the extrudate. This strain is the minimum strain across the cross-section of the steady state region of the extruded shaft (Fig. 7.66).

A typical force–displacement curve for the extrusion process is given in Fig. 7.67. After the process starts at point A, there is a steep increase in force due to the elastic deformation of the billet followed by its plastification and filling of the dies. In many processes a discontinuity in the curve between A and B can be observed, which is due to the upsetting of the billet in the container (billets are slightly smaller in diameter – about 0.1 to 0.2 mm – than the container). The die is filled until point C. There is a local maximum due to the contact between unlubricated sheared surfaces of the specimens and the die. After point D the deformation zone in the die-shoulder remains

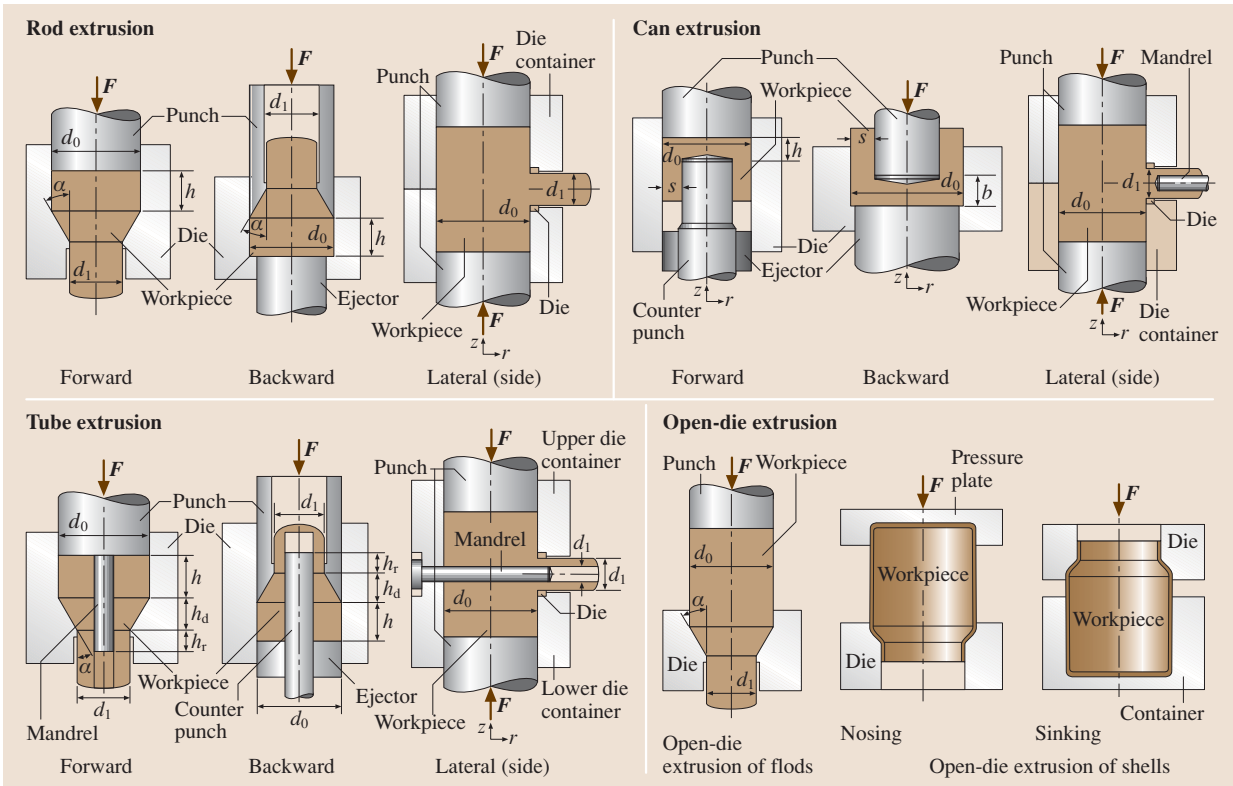


Fig. 7.64 Principles of typical extrusion processes

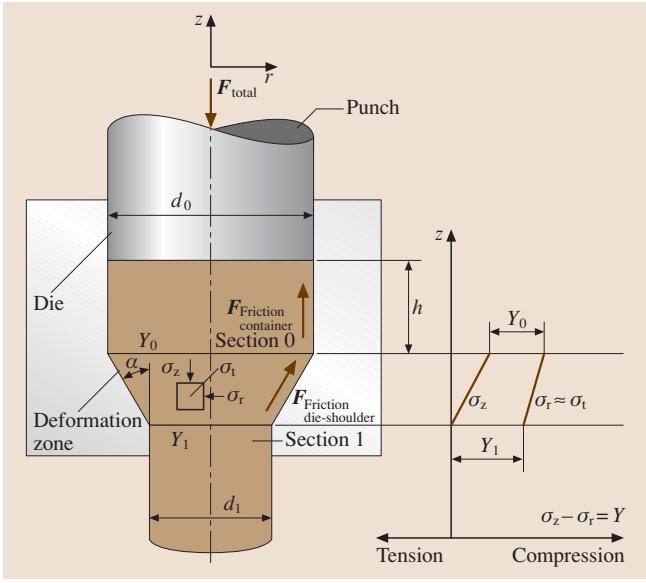


Fig. 7.65 Schematic sketch of forward (impact) extrusion of rods and stress states during extrusion (after [7.37])

steady so that there is only a decrease in the force due to the decrease in the friction forces in the die chamber.

The forming forces in extrusion of rods and tubes can be determined by the model suggested by Siebel. The total extrusion force consists in this case of the following parts

$$F_{\text{total}} = F_{\text{ideal}} + F_{\text{shear}} + F_{\text{friction container}} + F_{\text{friction die-shoulder}} + F_{\text{friction mandrel}} + F_{\text{friction mandrel-land}} \,. \quad (7.65)$$

The force components are described in Table 7.25. Here, σ_{fm} is the mean flow stress in the deformation zone given by

$$\sigma_{\text{fm}} = \frac{1}{\varphi_1 - \varphi_0} \int_{\varphi_0}^{\varphi_1} \sigma_{\text{f}}(\varphi)\,\mathrm{d}\varphi \,. \quad (7.66)$$

μ is the friction coefficient (between 0.04 to 0.08 for forward rod extrusion, 0.1 to 0.125 for tube extrusion), h is the billet length in the container, h_{r} is the length

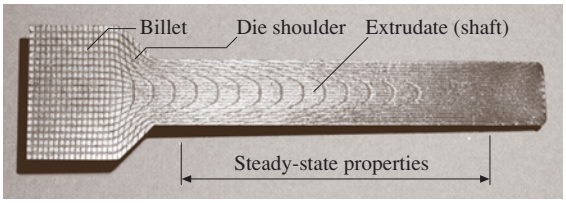


Fig. 7.66 Cross-section of a typical forward rod extrusion product with a deformation grid

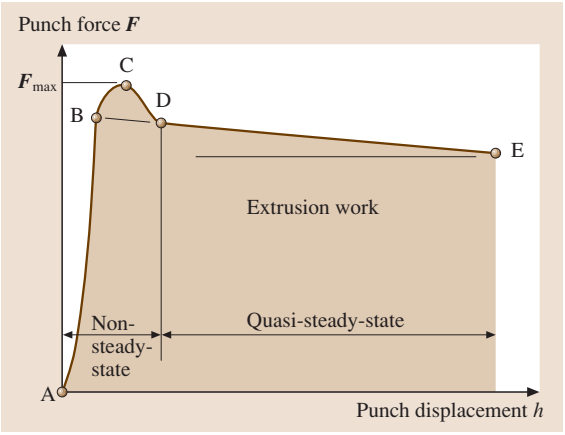


Fig. 7.67 A typical force displacement curve for forward rod extrusion

of the die exit land, and p_r is the pressure acting between the workpiece and the die (10–12 N/mm^2). For backward extrusion and the open-die container the friction is zero. For extrusion of profiles the cone angle $2\alpha = 180°$ usually. In this case, a dead-zone of workpiece material builds in the die so that the effective cone angle is $2\alpha = 90°$. This effective angle must be used in all force terms and the coefficient of Coulomb friction must be taken as 0.5 in the die-shoulder.

The effect of the die cone angle on various force components is exhibited in Fig. 7.68. The ideal force and the container friction is independent of the cone angle. The shear force increases with the cone

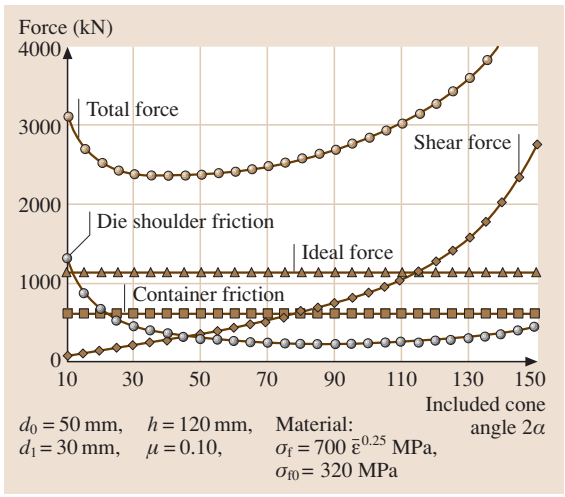


Fig. 7.68 Effect of various parameters on the extrusion force

angle, whereas the die-shoulder friction is minimum at $2\alpha = 90°$. Hence, the total force assumes a minimum value for the given process parameters around $2\alpha = 40°$.

The force for backward can extrusion can be computed by the equation of Dipper (Fig. 7.64)

$$F_{\text{punch}} = A_i \sigma_{f1} \left(1 + \frac{1}{3}\mu \frac{d_i}{b}\right) + \sigma_{f2} \left(1 + \frac{\mu + 0.5}{2} \frac{b}{s}\right) , \qquad (7.67)$$

where A_i is the punch cross-sectional area, μ the coefficient of friction (0.04 to 0.08), and the flow stresses

$$\sigma_{f1} = \sigma_f \left(\varphi = \ln\left[\frac{h_0}{b}\right]\right) ,$$
$$\sigma_{f2} = \sigma_f \left(\varphi = \ln\left[\frac{h_0}{b}\right]\left[1 + \frac{d_i}{8s}\right]\right) , \qquad (7.68)$$

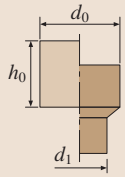
where h_0 is the initial height of the billet.

A simpler but effective equation for the punch load during backward can extrusion is given by Hoogen-

Table 7.25 Force computations for rod and tube extrusion (after [7.37])

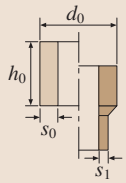
Process	F_{ideal}	F_{shear}	F_{Friction} (container)	F_{Friction} (die-shoulder)	F_{Friction} (mandrel-die)	F_{Friction} (mandrel-land)
Rod extrusion	$A_0 \sigma_{fm} \varphi$	$\frac{2}{3} \tan\alpha \sigma_{fm} A_0$	$\pi d_0 h \sigma_{f0} \mu$	$2\sigma_{fm} \varphi \mu A_0 / \sin 2\alpha$	0	0
Tube extrusion	$A_0 \sigma_{fm} \varphi$	$\frac{1}{2} \tan\alpha \sigma_{fm} A_0$	$\pi d_0 h \sigma_{f0} \mu$	$2\sigma_{fm} \varphi \mu A_0 / \sin 2\alpha$	$\sigma_{fm} \varphi \mu A_1 / \tan\alpha$	$\pi d_2 h_r p_r \mu$

a) Forward rod extrusion



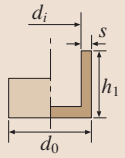
	$d_{0,\max}$ [a] (mm)	$\left(\frac{A_0}{A_1}\right)_{\max}$	$\varepsilon_{A,\max}$	$\left(\frac{h_0}{d_0}\right)_{\max}$ [a]
Aluminum (Al99.5), lead, zink	500–300	50	0.98	[c]
Copper (E-Cu)	350–200	5	0.80	[c]
Brass (Ms 63 to Ms 72)	300–150	4	0.75	[c]
Easy deformable steel (C10, C15, …)	250–125	3.3	0.70	8 – 3
Less easy deform. steel (Ck35, 16MnCr5)	220–110	2.2	0.55	6 – 2.5
Difficult to deform. steel (C45, 34Cr04)	200–100	2.0	0.50	4 – 2

b) Forward tube extrusion



	$d_{0,\max}$ [a] (mm)	$\left(\frac{A_0}{A_1}\right)_{\max}$	$\varepsilon_{A,\max}$	$\left(\frac{s_0}{s_1}\right)$	$\left(\frac{h_0}{d_0}\right)_{\max}$ [a]
Aluminum (Al99.5), lead, zink	500–300	50	0.98	⩾0.05	[c]
Copper (E-Cu)	350–200	5	0.80	⩾0.15	[c]
Brass (Ms 63 to Ms 72)	300–150	4	0.75	⩾0.20	[c]
Easy deformable steel (C10, C15, …)	250–125	3.3	0.70	⩾0.20	8 – 3
Less easy deform. steel (Ck35, 16MnCr5)	220–110	2.2	0.55	⩾0.30	6 – 2.5
Difficult to deform. steel (C45, 34Cr04)	200–100	2.0	0.50	⩾0.40	4 – 2

c) Backward can extrusion



	$d_{0,\max}$ [b] (mm)	$\left(\frac{A_0}{A_1}\right)_{\max}$	$\varepsilon_{A,\max}$	$\frac{s_{\min}}{d_0}$	$\frac{s_{\max}}{d_0}$	$\left(\frac{h_1}{d_1}\right)_{\max}$
Aluminum (Al99.5), lead, zink	500–300	50	0.98	⩾0.01	⩽0.40	5
Copper (E-Cu)	350–200	5	0.75	⩾0.03	⩽0.35	3.5
Brass (Ms 63 to Ms 72)	300–150	3.3	0.70	⩾0.05	⩽0.30	2.5
Easy deformable steel (C10, C15, …)	250–125	2.9	0.65	⩾0.04	⩽0.30	2.5 – 3
Less easy deform. steel (Ck35,16MnCr5)	220–110	2.5	0.60	⩾0.07	⩽0.25	2
Difficult to deform. steel (C45, 34Cr04)	200–100	2.2	0.55	⩾0.09	⩽0.22	1.5

[a] From the smallest ideal equivalent strain ($\varphi = 0.4$) to the largest strain. The basis for the value of d_0 is a 25 MN force.
[b] From $s/d_0 = 0.15$ to s_{min}/d_0 [c] Values unknown

Fig. 7.69 Process limits for impact extrusion (after [7.37])

boom for ideal plastic materials

$$\frac{p}{\sigma_f} = \frac{2+\sqrt{3}}{\sqrt{3}} + \frac{1}{3}\sqrt{\frac{6s+d_i}{s}} \quad \text{early stage,}$$

$$\frac{p}{\sigma_f} = \frac{2+\sqrt{3}}{\sqrt{3}} + \frac{1}{2\sqrt{3}}\sqrt{\frac{s}{b}+\frac{b}{s}} \quad \text{late stage,} \qquad (7.69)$$

where p is the punch pressure and σ_f is the constant flow stress. The early stage is the quasi steady state process, whereas the late stage corresponds to the transient deformation for a can bottom of

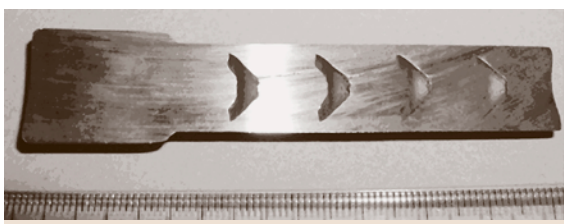
$$\frac{b}{s} < 5.5\frac{A_i}{A_e}\,, \qquad (7.70)$$

where A_e is the cross-sectional area of the billet. Other useful equations are given in [7.34].

For workpieces with nonaxial-symmetrical cross-sections such as bolts and nuts, the above equations can be applied replacing the diameters by equivalent diameters corresponding to the cross-section of the products.

Typical limits for the various extrusion processes are given in Fig. 7.69. The basic reason for limiting the process is the capacity of the forming presses and the tools. Another limiting factor is the formability of the material.

Typical failures in cold forged parts are internal and surface ductile cracks as well as galling. Examples of internal cracks are so-called Chevron cracks in extrusion



Material: 100Cr6 (not annealed) $2\alpha = 100°$ $\varphi = 0.25$

Fig. 7.70 Chevron cracks in cold extruded rods

Fig. 7.71 Typical process sequences of backward can extrusion

(Fig. 7.70). Galling is a type of adhesive wear. Adhesive wear occurs when the tools slide against the workpiece under high pressure after the breakdown of the lubricant. The heat generated due to friction causes microwelds to form between the sliding surfaces. Both the tools surface and the workpiece surface may be damaged.

Fig. 7.71 shows a typical sequence of processes for cold backward can extrusion. First, a slug is sheared off from a rod. Then these pieces are phosphated to carry the lubricant under high pressures. The lubricant in this case is soap. An upsetting process is performed in order to remove the irregular shear surface and supply a centering aid for the main extrusion process. Besides soap various other lubrications are possible (Table 7.26). Depending on the degree of deformation soap or MoS_2 or both with various pressure and fatty additives must be used.

A standard cold forging tool-setup is displayed in Fig. 7.72. It consists of reusable parts such as the base plates, pressure plates, intermediate plate, and the guides, as well as the product dependent interchangeable parts such as the punch, the die insert, and the shrink ring. The dies are exposed to radial, tangential,

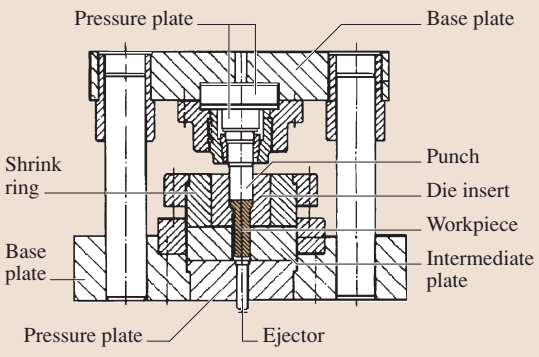


Fig. 7.72 Typical tool set-up for cold forward rod extrusion (after [7.44])

and equivalent stresses as shown in Fig. 7.73. With no precautions the maximum internal pressure can be

$$p_i \leq \frac{\text{yield strength of die material}}{2} . \qquad (7.71)$$

By imposing on the die a ring that has a smaller internal diameter than the outer diameter of the inner die, the sustainable internal pressure can be increased to

$$p_i \leq \text{yield strength of die material.} \qquad (7.72)$$

Drawing

Drawing can be broadly classified as drawing of solids parts such as rods, wires, or slabs, and drawing of hollow parts such as tubes or cans [7.45]. Figure 7.74a,b show typical wire drawing processes with the associated stress state in the deformation zone. Principally,

Table 7.26 Typical lubrication procedures (after [7.44])

Process	Deformation	Lubrication
Upsetting	Light	None
		Mi + EP + FA
	Severe	Ph + SP
Ironing and open-die extrusion	Light	Ph + Mi + EP + FA
	Severe	Ph + SP
Extrusion	Light	Ph + Mi + EP + FA
	Severe	Ph + SP
		Ph + MoS_2
		Ph + MoS_2 + SP

Mi = Mineral oil
SP = Soap
EP = Extreme pressure additive
Ph = Phosphate coating
FA = Fatty additives

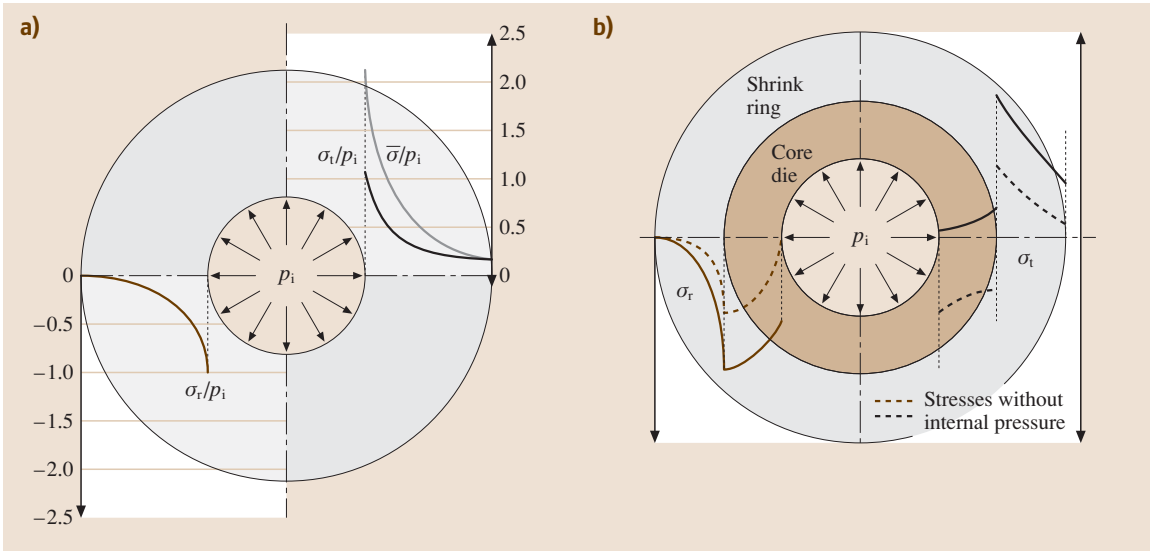


Fig. 7.73a,b Stresses in a die as a function of the internal forming pressure: **(a)** without a shrink ring, **(b)** with a shrink ring

the mechanics of deformation in drawing and extrusion are similar. The only difference is that in drawing the external force is applied on the exit side of die as a tension force as compared to extrusion where this force is on the entry side of the die and is compressive. Hence, the axial stresses in drawing processes are basically tensile compared to the compressive ones in extrusion. But both the radial and the circumferential stresses are compressive like in extrusion. In the case of the drawing of tubes, basically two broad categories are recognized: drawing with a mandrel or drawing without a mandrel. If a mandrel is used, this can be fixed, freely moving, or may have a guided move. One distinct group of tube drawing processes is the ironing process. This is often considered as a sheet metal forming process since basically the workpiece is a can obtained in deep-drawing. However, from the mechanics of deformation, this process is a typical bulk forming process in which the wall thickness of the can is reduced by drawing it through a die (Fig. 7.74b).

The drawing force can be determined by Siebel's equation

$$F_{\text{total}} = F_{\text{ideal}} + F_{\text{shear}} + F_{\text{friction die-shoulder}} + F_{\text{friction mandrel}} \,, \tag{7.73}$$

where the force components are given in Table 7.27. In the case of ironing, the punch force is given only by the first three components. However, if the force to be transmitted by the base of the can has to be determined, the force due to friction at the mandrel must be taken as negative.

The working window of the process of tube drawing is given in Fig. 7.75. There are basically three types of failure modes:

1. Rupture of the tube: The basic limitation of the process is that the tensile force must be applied through the formed tube portion. To prevent plastification leading to rupture

$$\sigma_z \leq \sigma_{f1} \,, \tag{7.74}$$

Table 7.27 Force computations for wire and tube drawing (after [7.37])

Process	F_{ideal}	F_{shear}	F_{Friction} (die-shoulder)	F_{Friction} (mandrel-die)
Wire and bar drawing	$A_1\sigma_{fm}\varphi$	$\frac{2}{3}\tan\alpha\sigma_{fm}A_1$	$2\sigma_{fm}\varphi\mu A_1/\sin 2\alpha$	0
Ironing and tube drawing over fixed mandrel	$A_1\sigma_{fm}\varphi$	$\frac{1}{2}\tan\alpha\sigma_{fm}A_1$	$2\sigma_{fm}\varphi\mu_s A_1/\sin 2\alpha$	$\pm\sigma_{fm}\varphi\mu_M A_1/\tan\alpha$

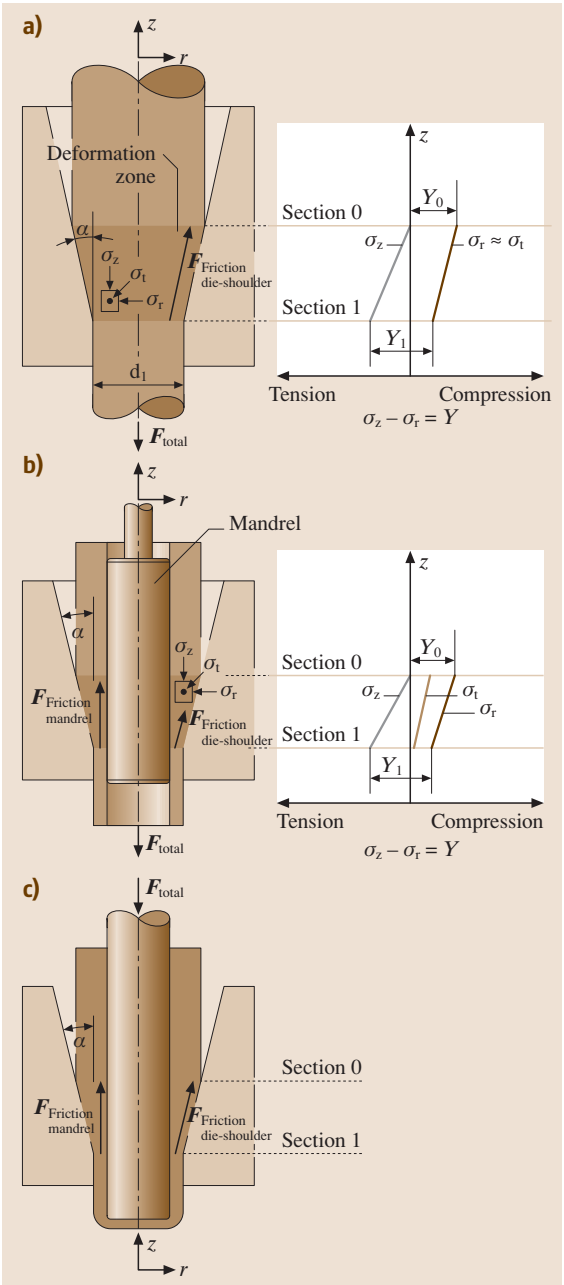


Fig. 7.74a–c Principles of typical drawing processes: (a) wire drawing, (b) tube drawing, (c) ironing

where σ_z is the uniform axial stress at the exit side and σ_{f1} is the mean flow stress at the same side. To ensure stability of the process usually the axial stress is taken less than 75% of the mean flow stress.

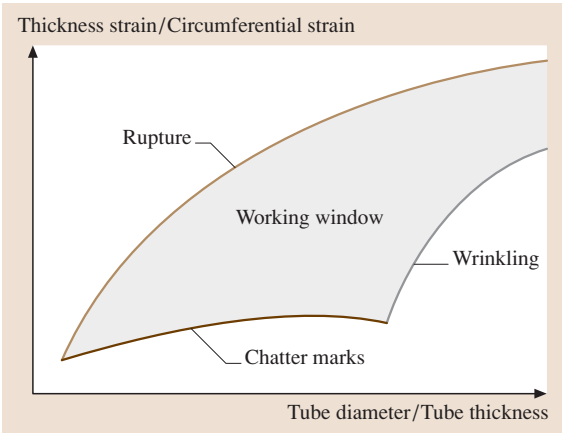


Fig. 7.75 Working window for tube drawing (after [7.37])

2. Chatter marks: Due to stick-slip effects the surface of the product shows marks that lower its quality. This defect can be avoided by increasing the stiffness of the system workpiece–tools.
3. Wrinkling: This failure type results from the compressive circumferential stresses in the tubes.

The maximum amount of area reductions for various wire materials during multiple die drawing is given in Table 7.28:

Wire drawing is basically a cold forming process. Residual stresses in the product are important both for service loading, as well as for subsequent processing of the material. Basically, harmful, large tensile residual stresses are left at the surface of the tube or rod after forming. Therefore, various precautions are suggested to reduce these high tensile residual stresses. One methodology is to use a double die design in which the second die has a very low area reduction. Figure 7.76 shows the effect of the after-reduction. Through a second low strain drawing die the high tensile stresses are reduced from about 550 to 250 MPa.

If the second area reduction is too small the plastification in the second die will be so low that the residual stress state will not change considerably; if, however, the area reduction is too large, this may cause a high plastic strain that may generate its own residual stress state that may be also not beneficial. Hence there is an optimum amount of area reduction and also a minimum distance between the two drawing dies.

Rolling

The forming process of passing a workpiece between rotating rolls is called rolling [7.30, 33]. According

Table 7.28 Allowable area reductions for wire drawing (after [7.38])

Material	Ultimate strength (MPa)	Initial wire diameter (mm)	Strain φ in each die	Total Strain	Number of draws
Steel	400	4–12	0.18–0.22	2.80–4.00	8–21
	1200	4–12	0.18–0.22	3.80–4.00	
	1200	0.5–2.5	0.12–0.15	1.20–1.50	
Cu-alloys	Cu (soft)	8–10	0.40–0.50	3.50–4.00	5–13
	250	1–3.5	0.18–0.20	2.00–2.00	
Al-alloys	Al (soft)	12–16	0.20–0.25	2.50–3.00	5–13
	80	1–3.5	0.15–0.20	1.50–2.00	

to the kinematics of the rolls and the workpiece motion rolling can be classified into three process families (Fig. 7.77). By the die geometry flat rolling (Fig. 7.78a) and profile rolling (Fig. 7.78b) processes are categorized. The process of reducing the thickness of a slab to produce a thinner and longer but only slightly wider product is referred to as flat rolling, whereas the forming by shaped rolls is known as profile rolling. Besides producing continuous products such as slabs, rolling is also used to form discrete parts such as bolts (Fig. 7.79) and rings.

Flat rolling is the basic forming process to produce plates and sheets. In the initial stages is conducted hot.

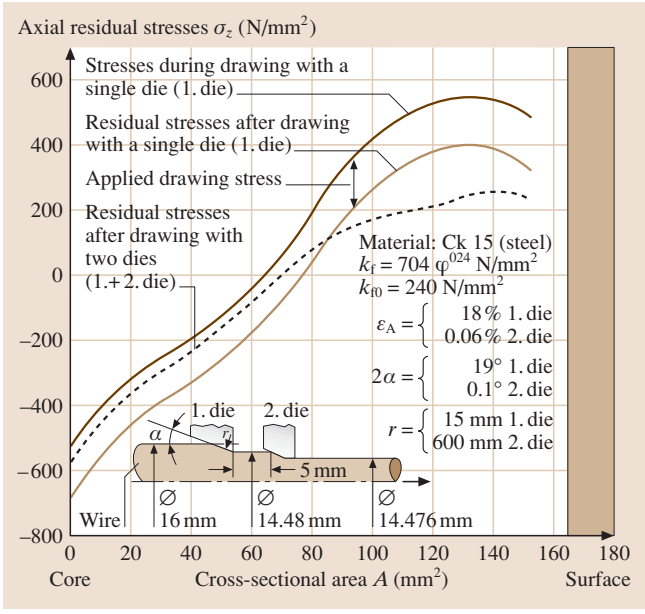


Fig. 7.76 Effect of second drawing die on the residual stresses (after [7.46])

For precision sheets the final passes are done cold. In both cases, the rolling process generates a typical texture in the products that is the source of anisotropic behavior of sheet products.

The variables of the rolling process are given in Fig. 7.80. It can be assumed that the width b of the workpiece is large enough such that during rolling no widening occurs. At a position of s_n a neutral plane exists. Before that neutral plane the workpiece material is slower than the rolls and after that it is faster. In order to ensure initial grasping of the workpiece by the rolls the contact angle α must fulfil

$$\alpha \le \mu \,, \tag{7.75}$$

where μ is the coefficient of Coulomb friction. The mean equivalent strain over the deformation zone is given by

$$\varphi = \ln\left(\frac{h_0}{h_1}\right) \,, \tag{7.76}$$

whereas the mean equivalent strain rate is

$$\dot{\varphi} = \varphi\omega\sqrt{\frac{R}{h_0 - h_1}} \,, \tag{7.77}$$

where ω is the rotational speed of the rolls in rad/s and R is the roll radius.

The vertical component of the pressure on the rolls can be approximated by [7.33]

$$\sigma(x) = \sigma_{fm}\left(1 + \frac{x^2/R + 2\mu x}{h_1 + x^2/R}\right) \quad \text{for} \quad 0 \le x \le s_n \,,$$

$$\sigma(x) = \sigma_{fm}\left(1 + \frac{2\mu R(s-x) - (s^2 - x^2)}{h_1 R + x^2}\right) \quad \text{for} \quad s_n \le x \le s \,. \tag{7.78}$$

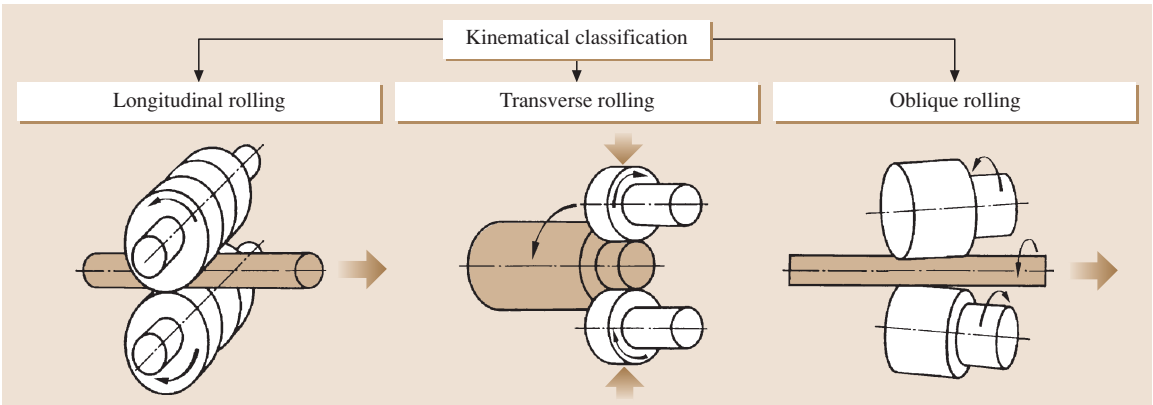


Fig. 7.77 Kinematical classification of rolling processes

The location of the neutral plane is given by

$$s_n = \frac{s}{2}\left(1 - \frac{s}{2\mu R}\right) , \qquad (7.79)$$

with

$$s = \sqrt{R(h_0 - h_1)} .$$

The mean flow stress is given in hot rolling for the mean equivalent strain rate and in cold rolling for the mean equivalent strain. The pressure on the rolls assumes a maximum value at the neutral plane (*friction hill*).

The rolling force F is found by

$$F = b \int_0^s \sigma(x)\,\mathrm{d}x . \qquad (7.80)$$

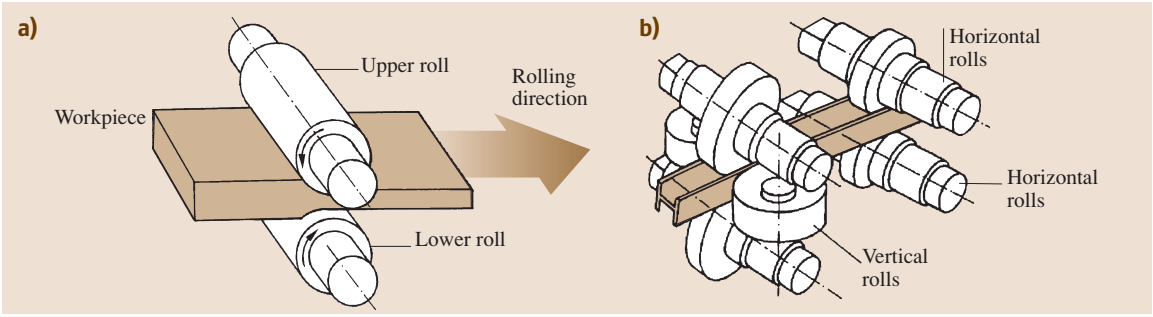


Fig. 7.78a,b Classification of rolling processes by die geometries: (a) Flat rolling, (b) profile rolling

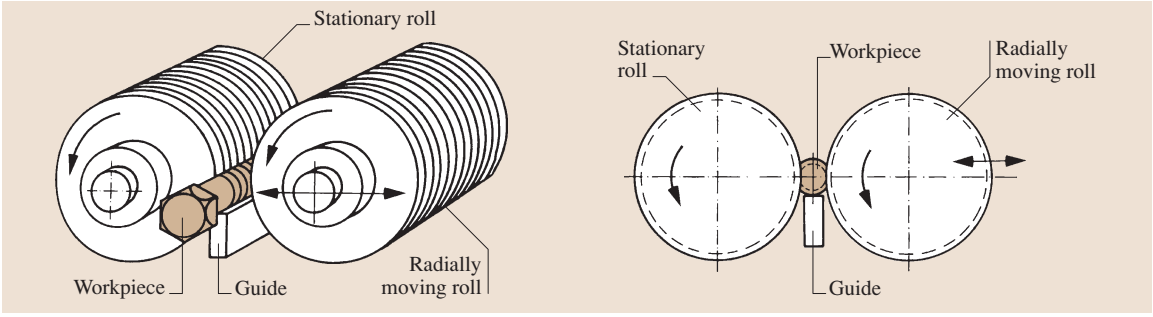


Fig. 7.79 Thread rolling of bolts

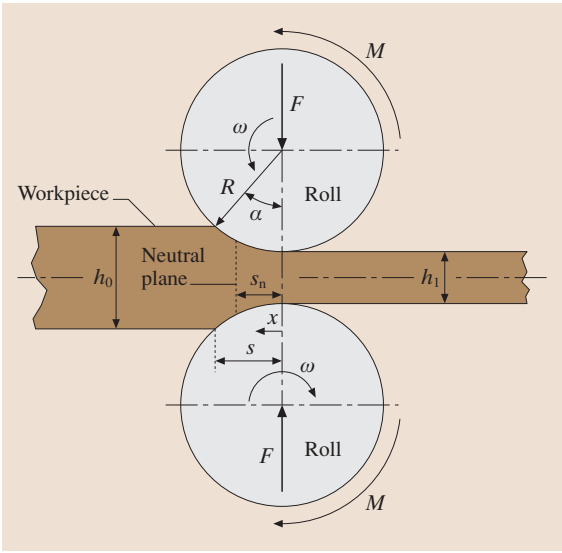


Fig. 7.80 Parameters of flat rolling

The rolling moment M of a single roll is given by

$$M = Fr\,, \tag{7.81}$$

where r is the moment arm given by [7.47]

$$r = 0.385s \quad \text{for} \quad \frac{R}{h_1} > 25\,,$$

$$r = \left[0.78 + 0.017\left(\frac{R}{h_1} - 0.163\sqrt{\frac{R}{h_1}}\right)\right] s \quad \text{for} \quad s \le 25\,. \tag{7.82}$$

The power necessary for the rolling process is

$$P = 2M\omega\,. \tag{7.83}$$

The rolls are exposed to a rather uniformly distributed forming load along their axis and hence behave like a simply supported beam and bend with maximum deflection at the center of the roll. To counter this deflection, which may lead to a rolled sheet with varying thickness, rolls are usually cambered, i. e. profiled rather than cylindrical. If the camber is insufficient, the thinner edges elongate plastically more than the center. Hence, in the rolled product compressive residual stresses are built at the edges and tensile residual stresses are built in the center parts (Fig. 7.81). This may lead also to centerline cracking, warping, or edge wrinkling. On the other hand, if the camber is too much, then the edge regions are elongated less plastically, leading to tensile residual

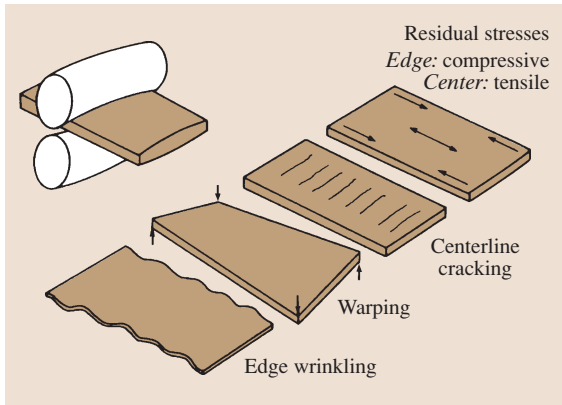


Fig. 7.81 Effects of insufficient camber of rolls (after [7.30])

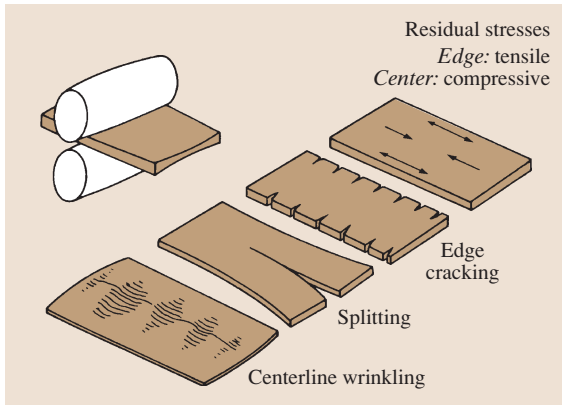


Fig. 7.82 Effects of excessive camber of rolls (after [7.30])

stresses at the edge region with edge cracking, centerline wrinkling, and splitting (Fig. 7.82). Roll bending can also be reduced by one or several back-up rolls.

Roll flatting is another limiting phenomenon of flat rolling. Due to the elastic deformations of rolls there is a minimum thickness of the end product [7.30]

$$(h_1)_{\min} = \frac{C\mu R}{E/(1-\nu^2)}(\sigma_{\mathrm{fm}} - \sigma_{\mathrm{t}})\,, \tag{7.84}$$

where C is a factor between 7 and 8, E is Young's modulus of the roll material, ν Poisson's ratio, σ_{fm} the mean flow stress for the deformation zone, and σ_{t} the front or back tension applied to the rolled sheet. Hence, the minimum thickness to be rolled can be decreased by decreasing friction, decreasing the roll diameter, increasing the stiffness of the rolls, decreasing the flow stress, and increasing the back and/or front tension.

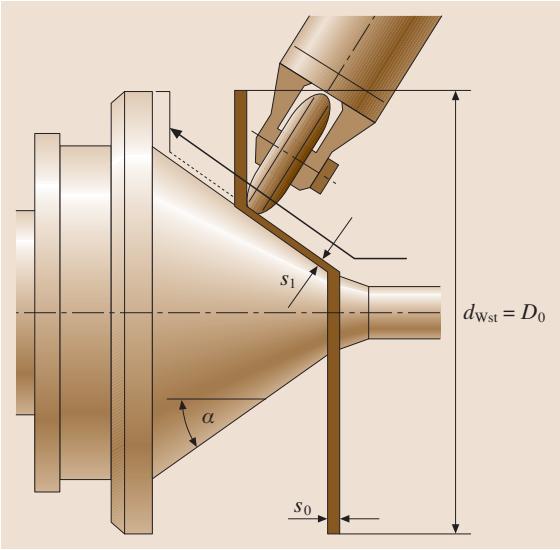


Fig. 7.83 Principle of shear forming process

Another process of rolling is the flow (shear) forming (turning) process (Fig. 7.83). In this process, a flat sheet is formed over a mandrel by means of a shear forming roll that is reducing its thickness but keeping its diameter constant. This process should not be mixed with sheet forming process spinning in which the sheet thickness is more or less constant but the diameter of the initial workpiece is reduced (Fig. 7.35). The final thickness of the workpiece is given by

$$s_1 = s_0 \sin\alpha \,. \tag{7.85}$$

If the final thickness is selected correctly the plastic deformation is confined only at the roll region, so that the rest of the workpiece remains stress-free. The key parameter of the process is the mandrel angle α. This angle must be less than 80° for localized plastic shear deformation and can be at least 12 to 18°. By a second pass this minimum angle can be reduced to even 8°.

The process of producing cylindrical parts is usually called flow forming whereas the forming of tapered parts is called shear forming.

7.2.5 Sheet Forming Processes

Membrane Theory

The approximate analysis of sheet forming processes can be performed by membrane theory [7.48]. Consider an axisymmetrical shell as shown in Fig. 7.84. The principal radii of curvature are in the hoop plane and in the meridian plane.

For thin plastically deforming shells, the bending moments are negligible and because of axial symmetry the hoop (σ_θ) and tangential (σ_ϕ) stresses are principal stresses (Fig. 7.85). The stress normal to the surface can be neglected, so that the resulting stress state is the one of plane stress. Friction forces are neglected. Only uniform pressure loads normal to the surface (although small enough with respect to the flow stress) and uniform edge tensions tangential to the surface are allowed.

It is furthermore assumed that work hardening is compensated by thinning of the sheet, so that the product of flow stress times current thickness is constant

$$\sigma_f t = T_f = \text{const.} \,, \tag{7.86}$$

where T_f is the so-called force resultant.

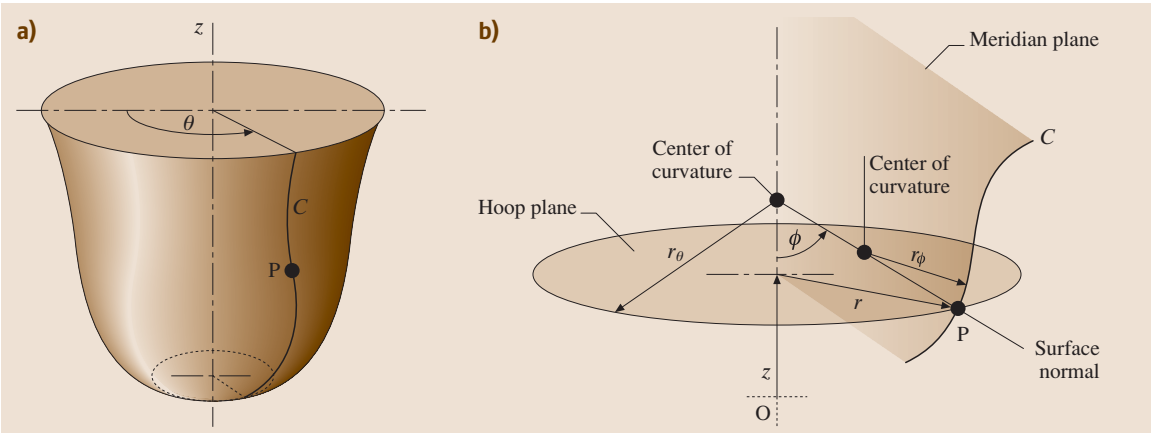


Fig. 7.84 (a) Axisymmetrical shell. (b) Radii of curvature r_θ and r_ϕ

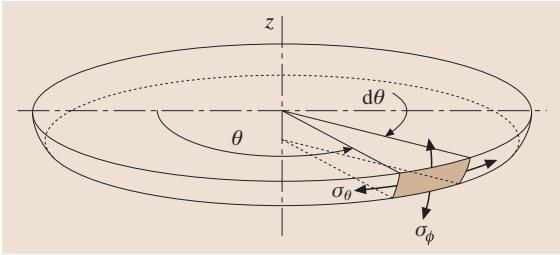
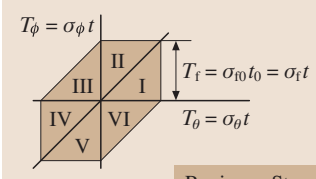


Fig. 7.85 Stress components in a shell



Region	Stress state	Flow condition
I	$T_\theta > T_\phi > 0$	$T_\theta = T_f$
II	$T_\phi > T_\theta > 0$	$T_\phi = T_f$
III	$T_\phi > 0 > T_\theta$	$T_\phi - T_\theta = T_f$
IV	$0 > T_\phi > T_\theta$	$T_\theta = -T_f$
V	$0 > T_\theta > T_\phi$	$T_\phi = -T_f$
VI	$T_\theta > 0 > T_\phi$	$T_\theta - T_\phi = T_f$

Fig. 7.86 Tresca flow criterion in terms of the force resultants

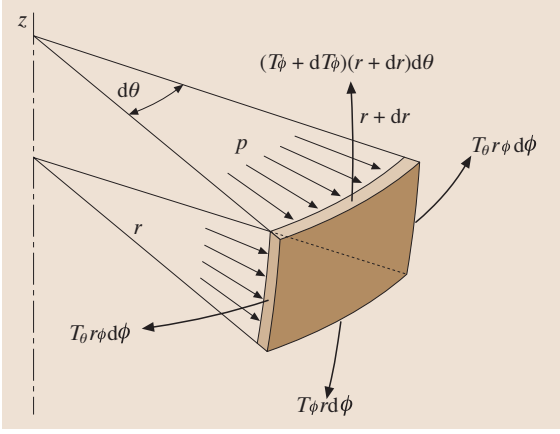


Fig. 7.87 Free body diagram of a typical shell element

Finally, the Tresca flow criterion is assumed to be valid as given in Fig. 7.86.

The static equilibrium of forces acting on a typical infinitesimal shell element (Fig. 7.87) supplies in the normal direction to the shell

$$p = \frac{T_\theta}{r_\theta} + \frac{T_\phi}{r_\phi} \tag{7.87}$$

and in the circumferential direction

$$\frac{dT_\phi}{dr} - \frac{T_\theta - T_\phi}{r} = 0\,. \tag{7.88}$$

Equations (7.87) and (7.88) together with the Tresca flow criterion (Fig. 7.86) constitute the framework of the membrane model to analyze sheet forming processes.

Plastic Anisotropy

Characterization of Anisotropy. Grains tend to assume preferred orientations during plastic deformation, since this deformation is achieved by slip or twinning. Preferred orientation of the grains will induce a direction dependent behavior of metals called anisotropy. The most obvious effect of anisotropy is observed in forming of sheets that are produced by rolling. Anisotropy of sheets results in earing at the rim of a deep drawn product (Fig. 7.109c).

Anisotropy in sheets is characterized by the Lankford parameter or anisotropy coefficient r. It is measured in the simple tension test (Fig. 7.88)

$$r = \frac{\varepsilon_b}{\varepsilon_t} = \frac{\ln \frac{b_1}{b_0}}{\ln \frac{t_1}{t_0}}\,, \tag{7.89}$$

where ε_b and ε_t are the true strains in width and thickness directions. For most metals the value of r changes

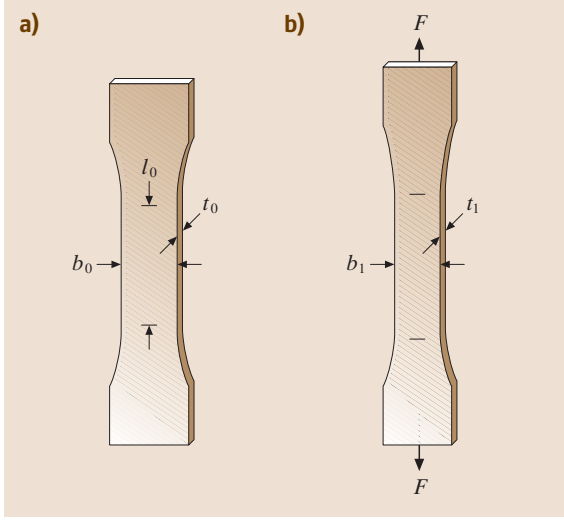


Fig. 7.88a,b Simple tension test of sheet metal: **(a)** undeformed, **(b)** deformed states

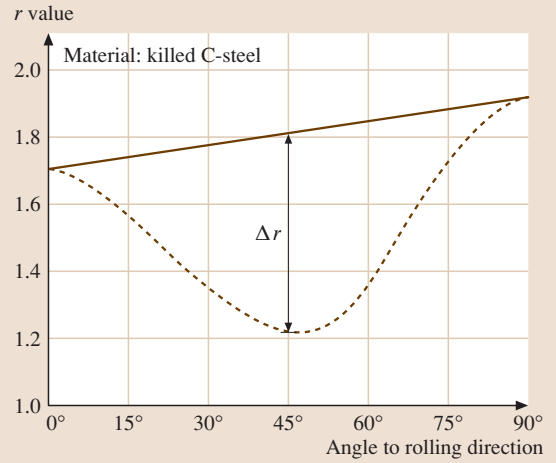


Fig. 7.89 Geometrical interpretation of the planar anisotropy coefficient

with the longitudinal strain, and by convention a longitudinal elongation of 20% is used for comparison purposes.

More importantly, the r-value changes with the orientation of the specimen with respect to the rolling direction. This variation is determined by three r-values obtained from specimens in the rolling direction, transverse to the rolling direction, and 45° to it. The average r-value, also called normal anisotropy r_n, is defined as

$$r_n = \frac{r_{0°} + 2r_{45°} + r_{90°}}{4} . \qquad (7.90)$$

Variation of the r-value over the plane of the sheet is described by Δr, the planar anisotropy coefficient

$$\Delta r = \frac{r_{0°} + r_{90°} - 2r_{45°}}{2} . \qquad (7.91)$$

Table 7.29 Anisotropy values for various materials

Material	r_n value	Δr
Deep drawing steels (DC01–DC07; cold rolled)	1.30 to 2.00	up to 0.70
Stainless steel	0.70 to 1.10	−0.25 to 0.20
TRIP steels	0.9	about −0.03
Aluminum alloys	0.60 to 0.80	−0.60 to −0.15
Copper	0.60 to 0.80	–
Brass	0.80 to 1.00	–
Zinc alloys	0.20 to 0.6	–
Titanium alloys	2.00 to 8.00	up to 4.00

The geometric interpretation of the planar anisotropy coefficient is given in Fig. 7.89.

Typical anisotropy values for various materials are tabulated in Table 7.29.

The Anisotropic Flow Condition. The oldest anisotropic flow condition was proposed by *Hill* [7.26]. Neglecting planar anisotropy and assuming plane stress states (which is usually justified for sheet forming processes), the criterion reads in the principal stress configuration as

$$\sigma_1^2 - \frac{2r}{r+1}\sigma_1\sigma_2 + \sigma_2^2 = \sigma_f^2 , \qquad (7.92)$$

where σ_1 and σ_2 are the in-plane principal stresses and σ_f is the flow stress. The effect of the normal anisotropy value r on the flow locus is shown in Fig. 7.90. For $r = 1$ the standard von Mises ellipse is obtained. For r-values larger than 1, the ellipse elongates along the major axis whereas it shrinks along the minor axis. Hence, in the case of biaxial tension or biaxial compression larger relative stresses are necessary to initiate plastic flow. In the second and forth quadrants plastic flow will oc-

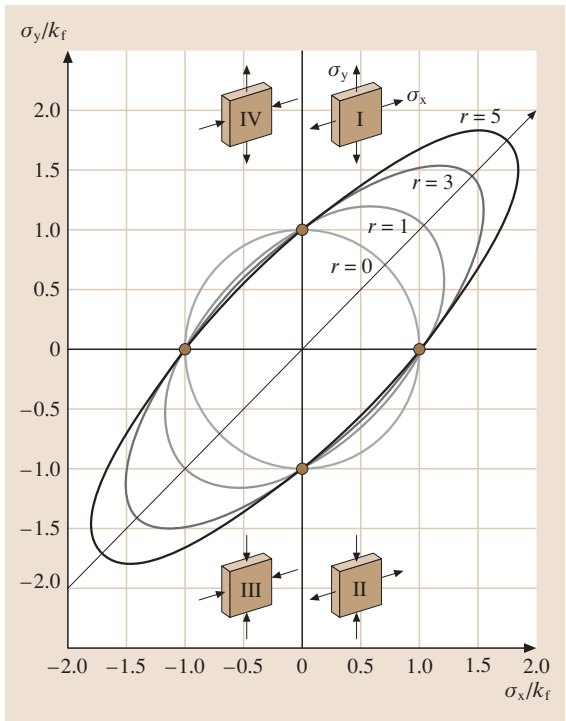


Fig. 7.90 Effect of the r value on the flow locus

Table 7.30 Various nonquadratic anisotropic flow criteria (flow planar isotropy) (after [7.36])

Flow criterion	Equation	Suggested parameter values
Hill 1979	$\lvert\sigma_1+\sigma_2\rvert^m+(2r+1)\lvert\sigma_1-\sigma_2\rvert^m=2(r+1)\sigma_f^m$	$1.3\le m\le 2.2$ (for $m=2$ criterion reduces to Hill 1948)
Hosford	$\sigma_1^a+r(\sigma_1-\sigma_2)^a+\sigma_2^a=(r+1)\sigma_f^a$	$a=6$ for ffc metals $a=8$ for bbc metals
Barlat 1989	$a\lvert k_1+k_2\rvert^M+a\lvert k_1-k_2\rvert^M+(2-a)\lvert 2k_2\rvert^M=2\sigma_f^M$ $k_1=\frac{\sigma_x+h\sigma_y}{2}$, $k_2=\left[\left(\frac{\sigma_x-h\sigma_y}{2}\right)^2+p^2\tau_{xy}^2\right]^{1/2}$	a, h, p and M are material parameters
Banabic–Barlat 2000	$(m\sigma_1+n\sigma_2)^{2k}+(p\sigma_1+q\sigma_2)^{2k}+(r\sigma_1+s\sigma_2)^{2k}=2\sigma_f^{2k}$ $m=\frac{\left(-\frac{\sigma_f}{\sigma_b}+4\right)}{3}$, $n=\frac{\left(\frac{\sigma_f}{\sigma_b}-4\right)}{3}$, $p=2\frac{\left(\frac{\sigma_f}{\sigma_b}-1\right)}{3}$, $q=\frac{\left(\frac{\sigma_f}{\sigma_b}+2\right)}{3}$, $r=\frac{\left(-\frac{\sigma_f}{\sigma_b}-2\right)}{3}$, $s=2\frac{\left(\frac{\sigma_f}{\sigma_b}+1\right)}{3}$	σ_f is the uniaxial flow stress and σ_b is the biaxial flow stress

cur for slightly lower relative stresses than the isotropic case.

Hill's criterion has proven itself successful for steel, however for aluminum particularly this criterion fails. Also, it is able only to predict two or four ears in deep-drawing, although a different number of ears is also observed for certain materials. Several criteria, also called nonquadratic flow criteria have been introduced for these reasons. Table 7.30 depicts a few of them. Most of them have been developed basically to model forming of aluminum sheets.

Formability

Basic Failure Modes. Sheet parts can fail under various modes (Fig. 7.91) There are three basic failure types:

1. Wrinkling caused by compressive stresses.
2. Rupture/tearing caused by exceeding the capability of the material, usually preceded by necking.
3. Surface defects.

Most of the tests to evaluate the formability of sheet metals in sheet forming operations and particularly in deep drawing and stretch forming are focusing on the fracture failure mode. Fracture usually follows a necking, i. e. a localized thinning of the metal. Therefore, necking is usually accepted as a failure mode as well.

Necking. Sheets neck under simple tension in two stages: diffuse necking and localized necking

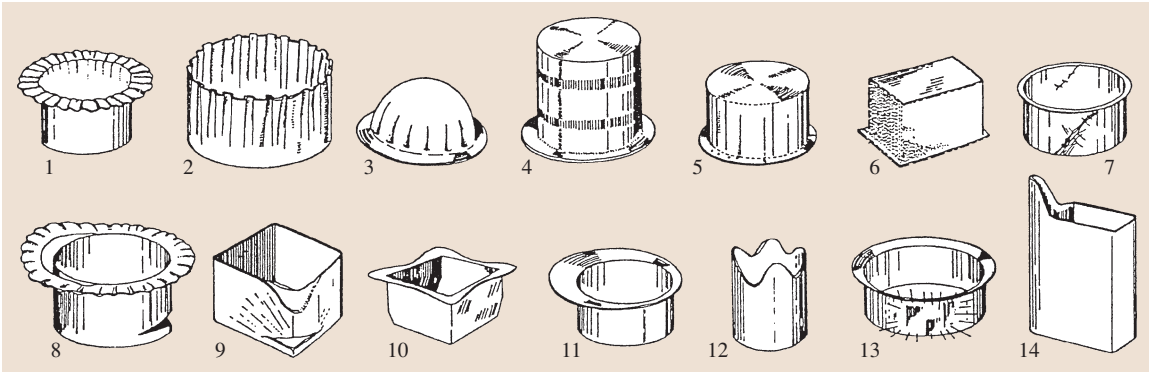


Fig. 7.91 Various failure modes in deep-drawing: 1 – flange wrinkling; 2 – wall wrinkling; 3 – part wrinkling; 4 – ring prints; 5 – traces; 6 – orange skin; 7 – Lüder's strips; 8 – bottom fracture; 9 – corner fracture; 10–12 – folding; 13,14 – corner folding (after [7.36])

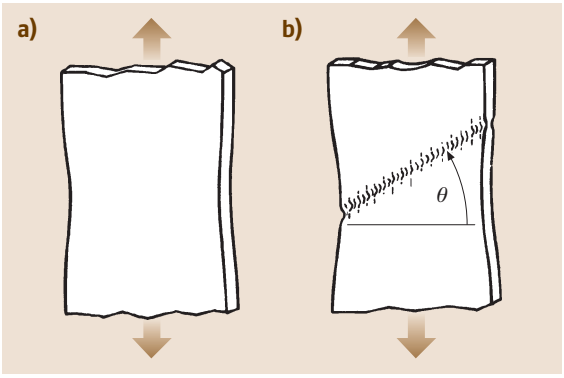


Fig. 7.92a,b Necking of sheet specimens under uniaxial tension: **(a)** Diffuse necking, **(b)** localized necking (after [7.30])

(Fig. 7.92); bulk specimens only exhibit diffuse necking. If the material's flow curve is represented by the Ludwik flow curve $\sigma_f = C \times \bar{\varepsilon}^n$, then it can be shown that

$$\text{for diffuse necking} \quad \varepsilon_1^{\text{necking}} = n \,,$$
$$\text{for localized necking} \quad \varepsilon_1^{\text{necking}} = 2n \,. \qquad (7.93)$$

If the stress state is not uniaxial, as is the usual case in most sheet forming processes, then the localized necking strain for an isotropic material is given as [7.30, 48]

$$\varepsilon_1^{\text{necking}} = \frac{n}{1 + \varepsilon_2/\varepsilon_1} \,. \qquad (7.94)$$

The localized necking angle with the tension axis is estimated as

$$\tan\theta = \frac{1}{\sqrt{-\varepsilon_2/\varepsilon_1}} \,, \qquad (7.95)$$

yielding an angle of 54° for simple tension of an isotropic specimen. For positive ratios of the two principal in-plane strains no localized neck is possible and only diffuse necking occurs.

Simulative Formability Tests. The Olsen or Erichsen test (Fig. 7.93) is the oldest test to evaluate formability of sheet metals. Both tests only differ slightly in the size of the tools. The principle is to stretch a sheet of metal by a hemispherical punch until fracture is observed. The punch depth in millimeters corresponds to the Erichson index (IE).

Another common test is the limiting dome height test by *Hecker* [7.49] (Fig. 7.94). Hecker's test is an improved Erichson test with larger tools and drawbeads increasing the measuring accuracy.

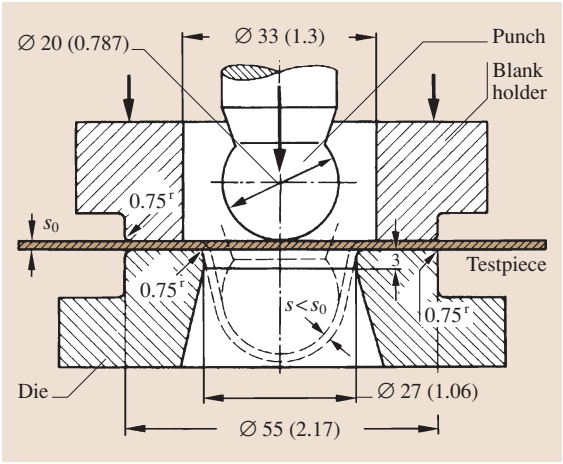


Fig. 7.93 Erichsen test (Dimensions in parentheses are in inches)

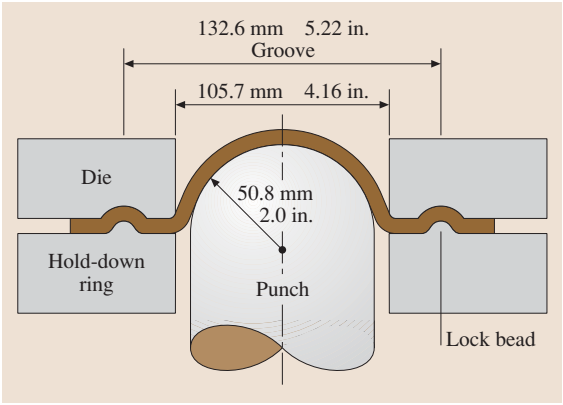


Fig. 7.94 Limiting dome height test by Hecker

Other tests such as Swift's cup test, the stretch–bend test, the wedge drawing test, the biaxial stretch test, the plane torsion test, hole expansion test etc. have been introduced in practice.

Forming Limit Diagrams. The forming limit diagram (FLD) consists of a curve in the principal in-plane strain space at which either necking starts or fracture is observed. It is basically a material property. Necking is localized in the tension–compression part and is expected to be diffuse in the tension–tension part (Fig. 7.95). The loading line on the left (dashed line) corresponds to a pure shear deformation, the central vertical line to plane strain tension, and the left dashed line to equibiaxial tension. Uniaxial simple tension is a loading line between the left dashed and the middle vertical lines.

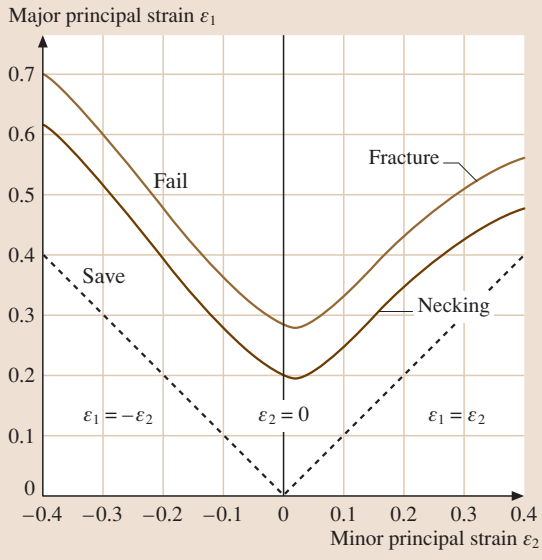


Fig. 7.95 Forming limit diagram for plane stress

FLD provided an efficient and practical method to assess the formability of a sheet product. It is applied in two different manners. *Experimentally*, a circular grid pattern is imposed on the sheet either mechanically, photochemically, or electrochemically. During forming the circles are deformed into ellipses. The principal strains are assumed to be along the major and minor axes of the ellipses. These minor and major radii of the ellipse are measured either manually or automatically by a digital camera. By comparing the measured local strains with the FLD, the range of safety for deep-drawing, the critical zones where necking and/or fracture are most likely to occur, the strain level and the favorable working conditions (blank-holder pressure, lubrication, placement of draw beads, etc.) can be determined, and hence the deep-drawing process can be improved. In this context a *severity index* ranging from 0 to 10 is introduced, indicating the distance of the measured strain state from the FLD-curve. *Numerically*, the analysis of the sheet forming process can be done, for instance, by finite element models a priori to the actual pressing and the computed strains can be compared with the respective FLD and the forming process can be assessed. This latter application is basic industrial practice today.

FLDs can be measured by various methods such as the uniaxial tensile test using specimens having various dimensions with and without notches (Fig. 7.96a), the hydraulic bulge test using elliptical dies (Fig. 7.96b), the

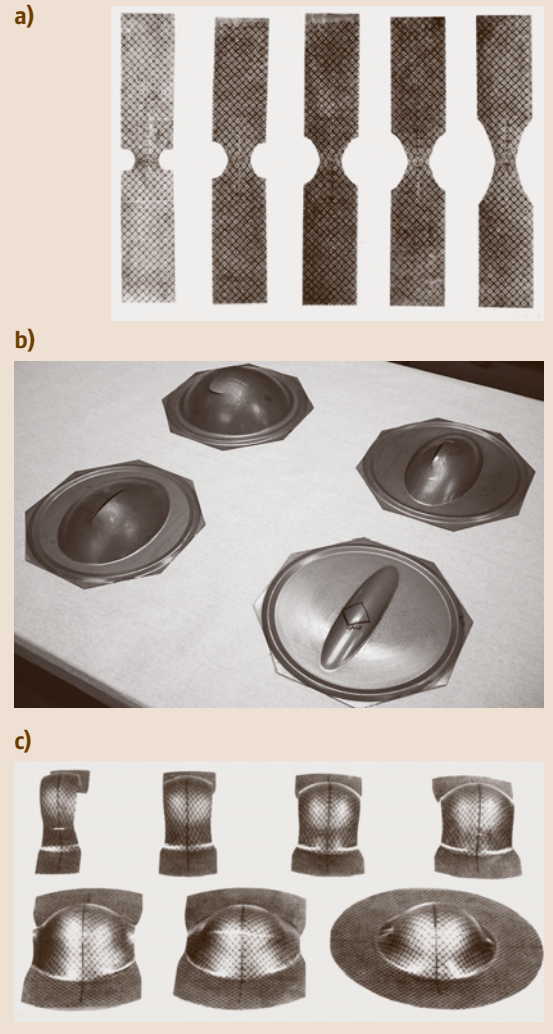


Fig. 7.96a–c Various methods of determining FLD: **(a)** uniaxial tensile test (courtesy V. Hasek), **(b)** hydraulic bulging test (courtesy N. Bay, DTU), **(c)** punch stretching test (courtesy V. Hasek)

punch stretching test using strips with various widths (Fig. 7.96c), the Nakamiza test (which is similar to the punch stretching test with the difference that the sheet is drawn instead of being stretched), the Hasek test (where circular specimens are used with various recesses), and the Marciniak test with hollow punches, etc.. Common to all tests is that numerous specimens or die geometries have to be used to simulate various principal strain ratios corresponding to various points on the FLD-curve.

The experimentally measured FLDs are not completely material properties. They depend on various other factors. It is known that the FLD-curve rises with increasing sheet thickness. Furthermore, it has been experimentally noticed that there is strain path dependence of the failure. Hence, if a tensile load path (meaning a strain path with positive slope on Fig. 7.95) is followed by a compressive load path (negative slope), the failure strain is lower than the one predicted by the FLD. On the other hand, if a compressive load path is followed by a tensile one, the failure strain is larger than the FLD strain. Also the grid size used in the experiments effects the measurements.

Bending

Bending Processes. Basic bending processes can be classified into two broad groups [7.50]: bending by linear tool motion (Fig. 7.97) and bending by rotary tool motion (Fig. 7.98).

In free bending [7.51] there is no contact between the workpiece and the die surface, whereas in die bending the sheet is bent between male and female dies. Free round bending is a continuous free bending process conducted in steps along the bending legs. In curling the workpiece is continuously bent by pushing it into a curved die. Bending by buckling is obtained by causing the workpiece to buckle normal to the applied force. The deformation region is limited either by local heating of the workpiece or clamping of the portion of the workpiece not to be deformed. In draw bending the sheet is shaped by being pulled through a die opening.

Roll bending is achieved by moving the workpiece through the gap created by three adjustable rolls. In roll forming the sheet is formed into a section between rolls with the form of the section. Folding is bending with a folding wing that folds the workpiece around the bending edge, whereas wiper bending or wiping is bending creating a full plastic section between a roller and a tool. The section is forced to conform to the roller as the tool is dragged around the roller. In roll straightening the roll axes can be normal or inclined to the bending plane. The workpiece is roll bent by corrugated rolls in the corrugating process.

The Mechanics of Bending. If a strip of width w and thickness t is bent by pure moments M at both ends, the strip assumes a curvature with radius r at its neutral axis (Fig. 7.99). The directions 1–3 are assumed as shown in the figure.

The bending strain for any axial fiber is given by

$$\varepsilon_1 = \ln\left(1 + \frac{y}{r}\right) , \tag{7.96}$$

where y is the distance of the fiber from the neutral axis having a radius of curvature r. This strain is positive at the upper fibers and negative at the lower fibers.

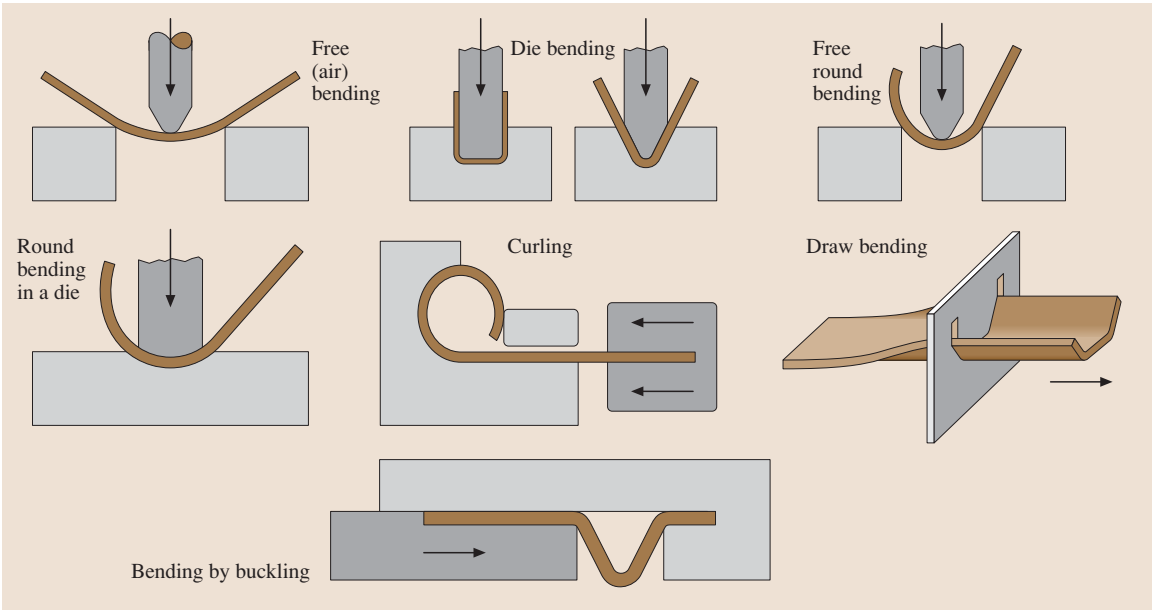


Fig. 7.97 Bending with linear tool motion

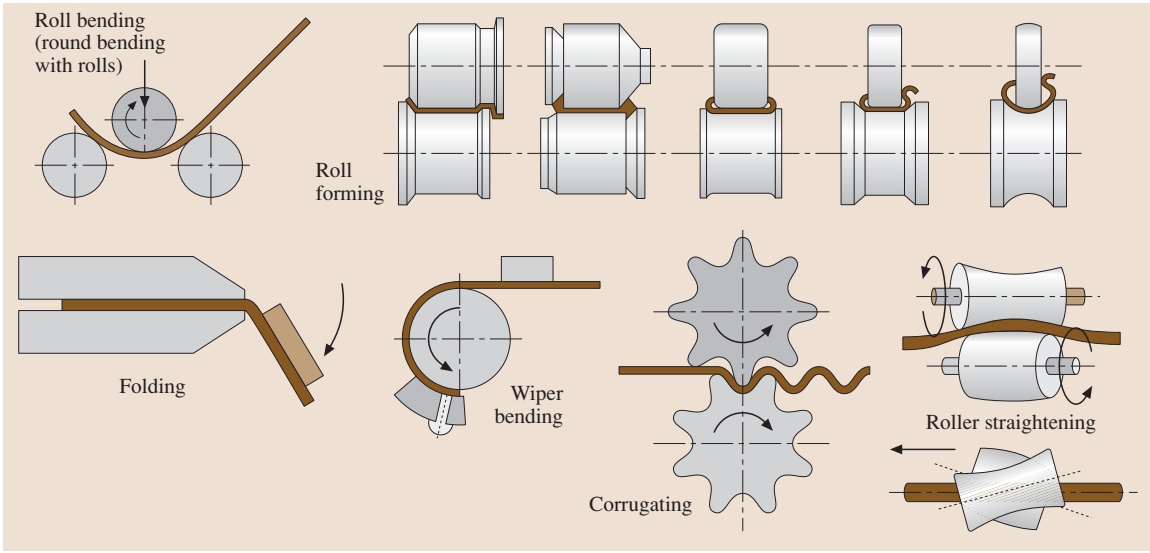


Fig. 7.98 Bending with rotary tool motion

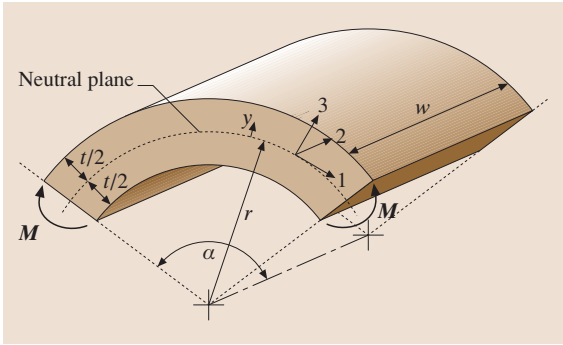


Fig. 7.99 Bending of a strip by pure moments

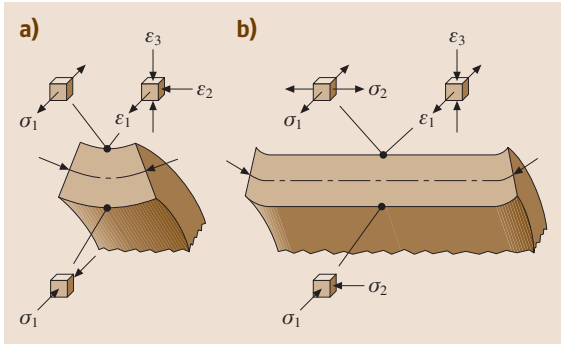


Fig. 7.100a,b Stress and strain states for bending of **(a)** narrow and **(b)** wide strips (after [7.52])

There are two extreme bending cases [7.52]: If the width of bent strip w is small as compared to its thickness t (Fig. 7.100a), then only the bending (axial) stress will develop, and hence the stress state is uniaxial with tensile stresses at the larger radius side and compressive stresses at the smaller radius side. The corresponding strain state is therefore

$$\varepsilon_1 = \varepsilon_2 = -\frac{1}{2}\varepsilon_1 \,. \tag{7.97}$$

Hence the width of the tension side of the strip will decrease and the width of the compressive side will increase. On the other hand, if the width of the bent strip is much larger than its thickness then in the middle regions of the strip a plane strain state will develop (Fig. 7.100b) with $\varepsilon_2 = 0$. If furthermore, the plane stress state is assumed with $\sigma_3 = 0$, it can be shown using the flow rule that $\sigma_2 = \sigma_1/2$. For this case of the plastic state, the flow stress and the equivalent strain is related to the axial components by

$$\varepsilon_1 = \sqrt{\frac{3}{2}}\bar{\varepsilon}\,,$$
$$\sigma_1 = \frac{2}{\sqrt{3}}\sigma_\mathrm{f}\,. \tag{7.98}$$

Even for a wide strip, at the ends an uniaxial stress state exists leading to a transverse curvature at both ends.

The elastic stress-strain relationship for the plane strain case is given by the generalized Hooke's law as

$$\sigma_1 = \frac{E}{1-\nu^2}\varepsilon_1\,, \tag{7.99}$$

where E is Young's modulus and ν is Poisson's ratio. Introducing the plane strain Young's modulus E', the elastic stress–strain relation gets

$$\sigma_1 = E'\varepsilon_1 \quad \text{with} \quad E' = \frac{E}{1-\nu^2} \,. \tag{7.100}$$

Figure 7.101a shows the axial strain distribution over the thickness of a strip. The true strain distribution is nearly linear. If the strains are large enough the material will plastify in the outer regions. Assuming that the material is perfectly plastic and the strip is wide, in the plastic regions the axial stress will be equal to the flow stress for plane strain as given by (7.98) (Fig. 7.101b). Assuming that the whole cross-section plastifies, the bending moment is given by

$$M = \frac{1}{2\sqrt{3}} W t^2 \sigma_\mathrm{f} \,. \tag{7.101}$$

After the moment is removed the bent part will spring back due to the elastic energy stored during bending (Fig. 7.102). The new radius of curvature will now be r'

$$r' = \left(\frac{1}{r} - \frac{2\sqrt{3}}{t} \frac{\sigma_\mathrm{f}}{E'} \right)^{-1} \,. \tag{7.102}$$

The residual stresses after unloading are given for the plastified region by (Fig. 7.101c)

$$\sigma_1^{\text{residual}} = \frac{2}{\sqrt{3}} \sigma_\mathrm{f} \left(1 - \frac{3y}{t} \right) \,. \tag{7.103}$$

The springback factor is defined by

$$\text{springback factor} \equiv \frac{r}{r'} \tag{7.104}$$

and is given for various materials and two relative bending radii in Table 7.31.

The basic failure mode in bending is fracture at the tensile side of the sheet. The smallest radius a sheet can be bent without failure is considered as its bendability. The bendability is highest if the bending axis is transverse to the rolling direction.

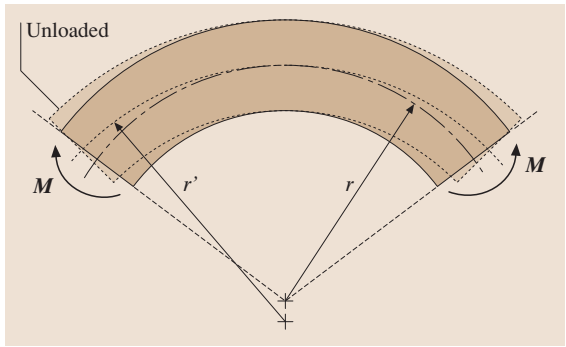


Fig. 7.102 Springback after unloading

Table 7.31 Springback factor (after [7.53])

Material	Springback factor $r'/t = 1$	Springback factor $r'/t = 10$
St 0-24, St 1-24	0.990	0.970
St 2-24, St 12	0.990	0.970
St 3-24, St 13	0.985	0.970
St 4-24, St 14	0.985	0.960
Stainless austentic steels	0.960	0.920
High temperature ferritic steels	0.990	0.970
High temperature austentic steels	0.990	0.970
Nickel w	0.990	0.960
Al99.5F7	0.990	0.980
AlMg1F13	0.980	0.900
AlMgMnF18	0.985	0.935
AlCuMg2F43	0.910	0.650
AlZnMgCu1.5F49	0.935	0.850

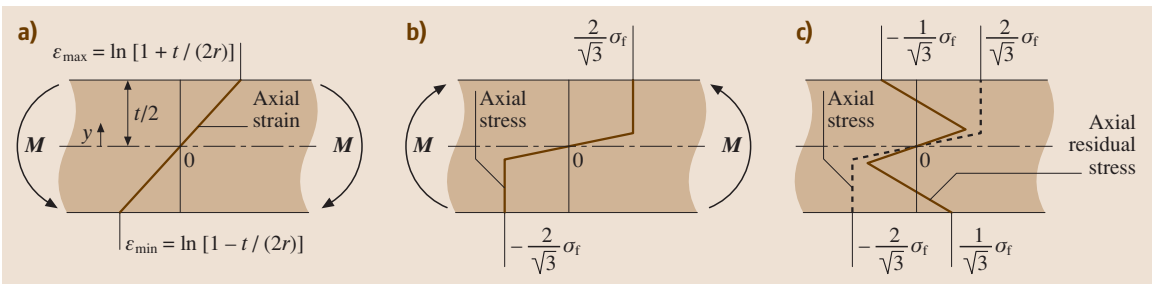


Fig. 7.101 (a) Strains, (b) loading stresses, and (c) residual stresses for a perfectly plastic material during plane strain bending

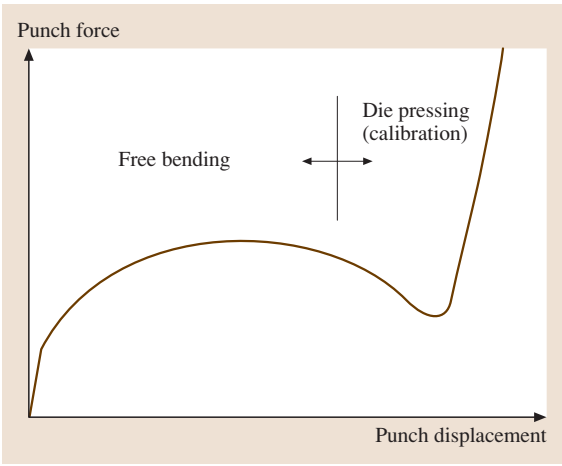


Fig. 7.103 Force–displacement curve in die bending

A typical force displacement curve for die bending (Fig. 7.97) is given in Fig. 7.103. The force increases during free bending and starts to drop as the sheet overbends. The steep increase at the end of process is as the bent sheet is pressed towards the die by the punch for calibrating the part.

Stretch Forming and Deep Drawing

Industrial sheet forming processes such as the manufacturing of automobile panels involve both stretch forming as well as deep-drawing processes. Stretch forming and deep drawing are similar processes in the sense that they shape a sheet using a rigid punch and a rigid die. On the other hand, the differ fundamentally in the sense that in stretching the sheet is fixed at its circumference, and hence the whole in-plane deformation is achieved by thinning of the sheet, whereas in deep drawing the sheet is allowed to be drawn at its circumference so that the shape change of the sheet is achieved under rather unchanged sheet thickness.

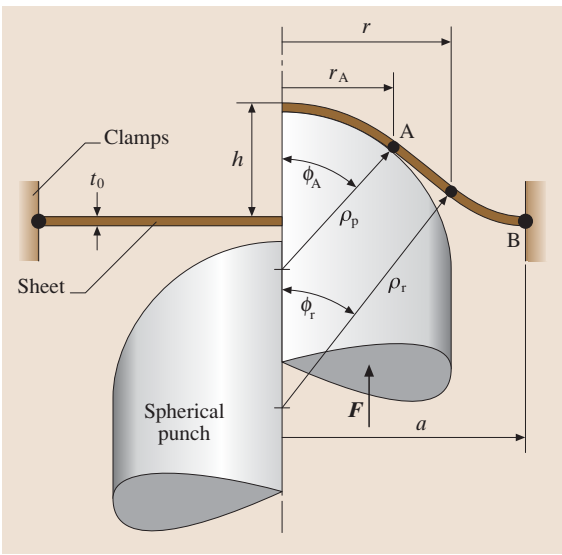


Fig. 7.104 Stretch forming with a spherical punch (after [7.52])

Stretch Forming. A circular sheet blank with diameter $2a$ being clamped firmly at its rim is stretched by a spherical punch with radius ρ_p (Fig. 7.104).

Using the membrane equation (7.87) it can be shown that the pressure between the punch and the sheet is

$$p_c = 2\frac{T_f}{\rho_p}\,, \tag{7.105}$$

and the punch force for a given contact angle ϕ_A is found as

$$F = T_f(2\pi\rho_p)\sin^2\phi_A\,. \tag{7.106}$$

Recall that the force resultant T_f is assumed constant and is therefore given as

$$T_f = \sigma_{f0}t_0 = \text{const.}\,, \tag{7.107}$$

where t_0 is the initial sheet thickness and σ_{f0} the initial flow stress.

The shape of the stretched sheet is given by the variable radius of curvature of the unsupported sheet region

$$\rho_r = \frac{r^2}{r_A\sin\phi_A}\,. \tag{7.108}$$

Deep Drawing. In the deep-drawing process an initially flat sheet (the blank) is formed by the rigid punch into the die cavity (Fig. 7.105). In order to prevent wrinkling of the blank a blank holder is usually used. Since the material in the blank is drawn in, the circumferential stresses are compressive here. The strain state at the rim is given by

$$\begin{aligned}
\varepsilon_\theta &= \ln\left(\frac{r_0'}{r_0}\right) \quad \text{(circumferential strain)}\,,\\
\varepsilon_t &= -\frac{1}{2}\varepsilon_\theta \quad \text{(thickness strain)}\,,\\
\varepsilon_r &= -\frac{1}{2}\varepsilon_\theta \quad \text{(radial strain)}\,.
\end{aligned} \tag{7.109}$$

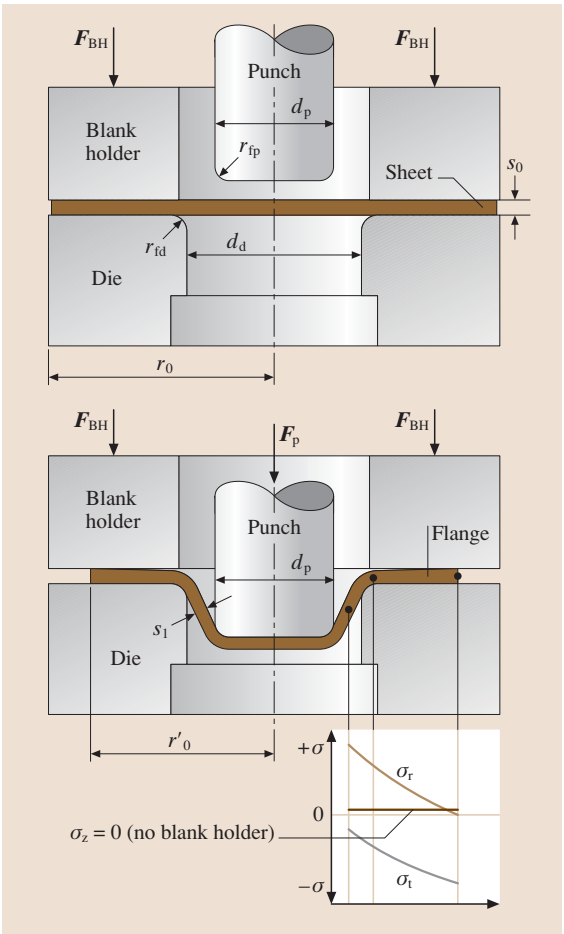


Fig. 7.105 Principle of deep drawing

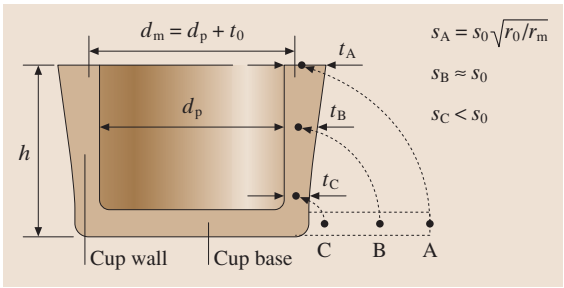


Fig. 7.106 Deep drawn cup (after [7.50])

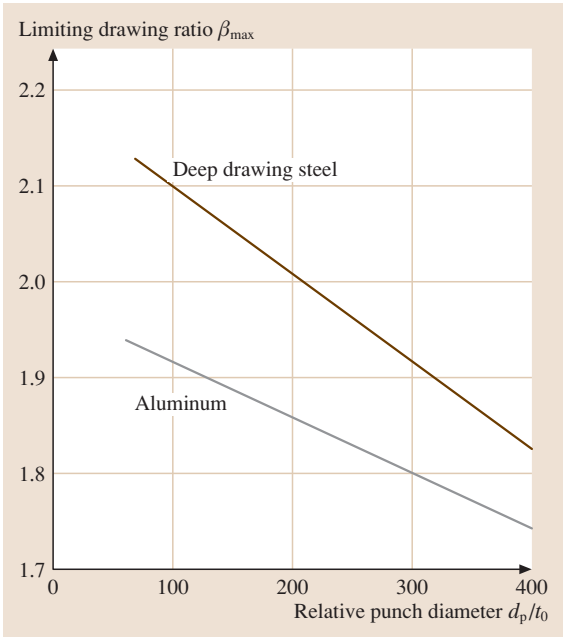


Fig. 7.107 Effect of relative punch diameter on the limiting drawing ratio

The wall thicknesses of the drawn cup are given in Fig. 7.106. Assuming that the mean thickness of the cup wall is equal to the initial sheet thickness, the drawn cup height is given as

$$h \approx \frac{\left(r_0^2 - r_m^2\right)}{2r_m} \tag{7.110}$$

with an error of ±5%.

The limiting drawing ratio β_{max} is an indicator of the drawability of a material and is defined as

$$\beta_{max} = \frac{d_{0,max}}{d_m} , \tag{7.111}$$

where $d_{0,max}$ is the largest blank diameter that can be drawn into a cup with mean diameter of d_m without failure. By membrane theory the ideal limiting drawing ratio is

$$\beta_{max} \leq e \approx 2.72 . \tag{7.112}$$

In reality, the limiting drawing ratios are 1.8 to 2.0 for aluminum sheets and 1.9 to 2.2 for steel sheets (for $d_0 = 100$ mm). This is because various process parameters are not considered in the simplifying membrane theory. One such factor is the dimension of the blank or equivalently the punch. As the relative punch diameter increases the limiting drawing ratio decreases (Fig. 7.107). Furthermore, as the normal anisotropy r_n increases the limiting drawing ratio increases; however, as the planar anisotropy Δr increases β_{max} decreases. Similarly, as the hardening

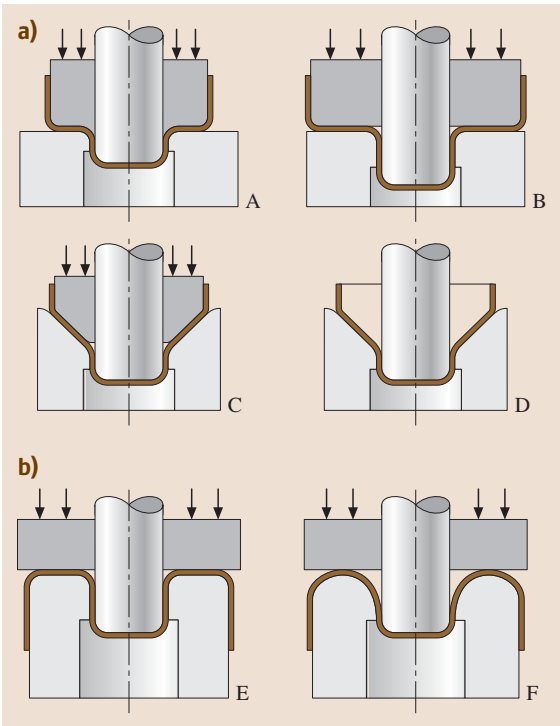


Fig. 7.108a,b Redrawing process to increase the overall drawing ratio: **(a)** direct redrawing, **(b)** inverse redrawing

exponent n increases β_{max} increases, but as friction between the blank and the die increases, β_{max} decreases.

To increase the drawing ratio cups can be drawn successively. This can be achieved by direct redrawing or reverse redrawing (Fig. 7.108). In direct redrawing the outer surface of the initially drawn cup remains the outer surface in the redrawn cup, whereas the same surface gets the inner surface in the reversely redrawn cup. Redrawing can be done either with annealing between the draws or without annealing. By the redrawing technique the total drawing ratio achieved without annealing between the draws is up to 6.5 for deep-drawing steel. Hereby, the first draws is conducted at a drawing ratio of 2.0 and all the following draws at a drawing ratio of 1.3.

The height of a drawn cup can be increased by reducing the wall thickness of the drawn cup by ironing (Fig. 7.74c). This process is also used to make the wall thickness of a cup uniform. This is virtually a drawing process and must be considered bulk forming.

Typical failure modes in deep drawing are given in Fig. 7.109. Tearing occurs if the drawing force can not be transmitted by the cup bottom; this is the failure type controlling the limiting drawing ratio. The limiting drawing force is

$$F \leq a\mathrm{UTS}\pi d_p t_0 \,, \qquad (7.113)$$

where a is a correction factor (for steel 1.05 to 1.55, for aluminum 0.99 to 1.22, for brass 0.92 to 1.27) and UTS is the ultimate tensile strength of the sheet material. Wrinkling is caused by compressive stresses that occur in the flange of the workpiece. If the clearance between the punch and the die is large, wrinkles can also occur in the cup wall. If the sheet thickness is large enough ($d_0/t_0 < 25$ to 40) the sheet has sufficient buckling resistance. For thinner sheets, wrinkles in the flange are prevented by a blank-holder. It is suggested to use a blank-holder pressure of 1 to 2% of the flow stress of the workpiece material.

The working window for the deep-drawing process is sketched in Fig. 7.110. The working window can be enlarged by reducing the flow stress of the workpiece in the flange region, by reducing friction in the flange and at the die radius, by increasing the strength of the material at the cup base/wall transition, and by increasing friction at the cup base region.

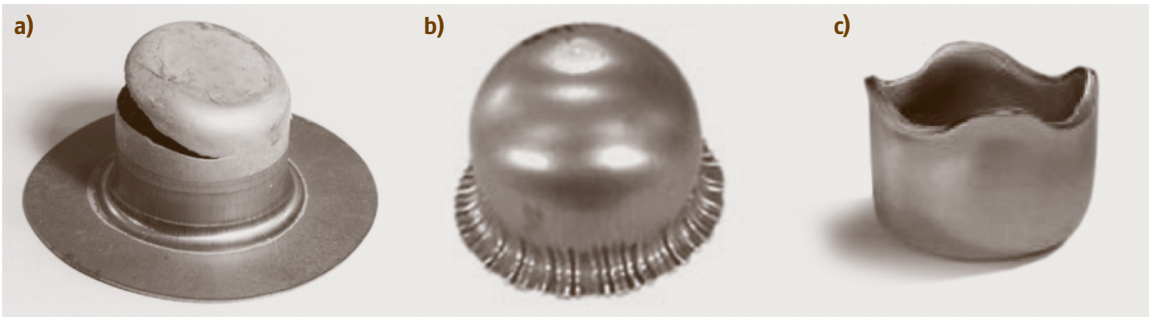


Fig. 7.109a–c Typical failure modes in deep drawing: **(a)** Tearing, **(b)** wrinkling (courtesy T. Altan, ERC), **(c)** earing (courtesy Hydro Aluminium Deutschland GmbH)

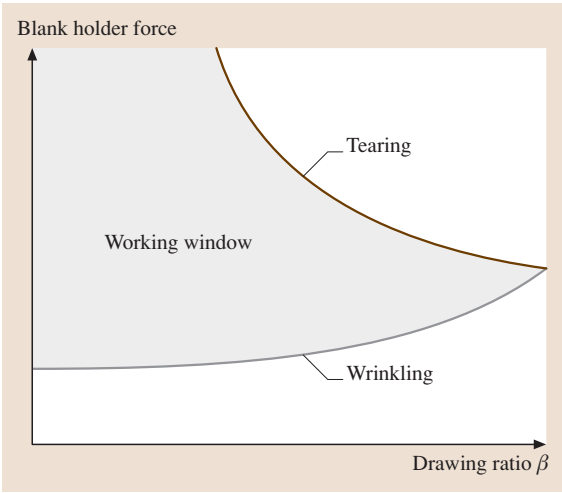


Fig. 7.110 Working window in deep drawing (after [7.52])

The drawing force during deep-drawing (see Fig. 7.111) can be estimated by

$$F = 2\pi r_m \left\{ \left[\underbrace{\sigma_{fm,flange} t_0 \ln\left(\frac{r'_0}{r_m}\right)}_{\text{ideal force}} + \underbrace{\frac{\mu F_{BH}}{\pi r'_0}}_{\text{flange friction}} + \underbrace{\frac{\sigma_{fm,die\text{-}ring} t_0^2}{2r_{fd} + t_0}}_{\text{bending/unbending}} \right] \underbrace{e^{\mu\phi}}_{\text{die-ring friction}} \right\} \sin\alpha \,, \quad (7.114)$$

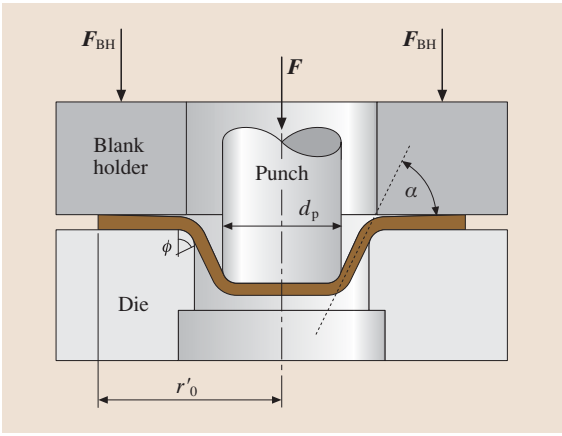


Fig. 7.111 Deep drawing parameters

The typical force–displacement in deep drawing shows a maximum as the blank radius reduces to around 77% of the initial value (Fig. 7.112). This is the result of the decreasing flange volume and the increasing flow stress in the flange as the punch advances. If the clearance between the punch and the die ring is smaller than the increased sheet thickness towards the rim of the cup, there will be an additional increase in the force due to ironing after the force maximum is reached.

Forming of complex three-dimensional sheets is a combination of stretch forming and deep drawing. The critical issue is to control the material flow within the sheet part during forming. This is achieved basically by four methods.

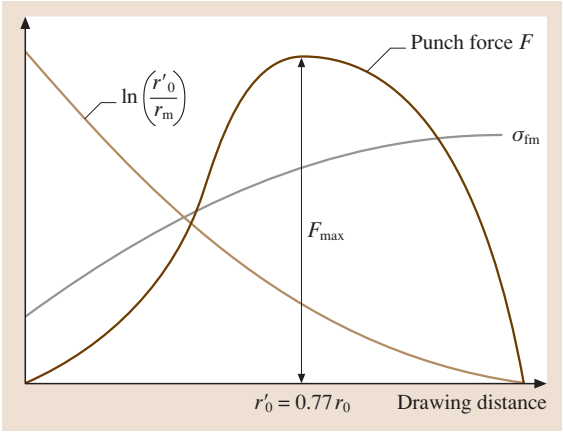


Fig. 7.112 Typical force–displacement curve for deep drawing

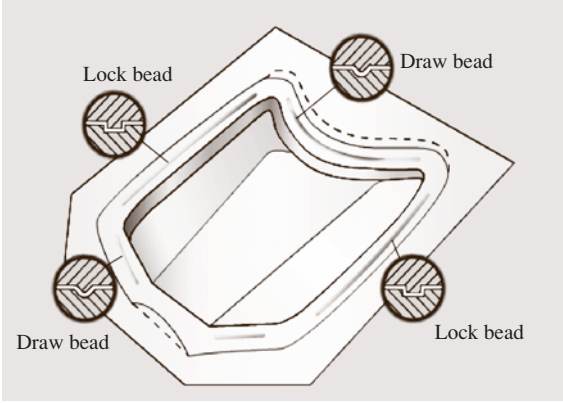


Fig. 7.113 Deep drawing of complex parts using draw beads and lock beads

1. Designing the blank geometry appropriately.
2. Using draw beads or lock beads (Fig. 7.113) that prevent partially or completely the draw-in of the material.
3. Regulating the blank-holder force. Advanced deep drawing dies have segmented blank-holders to adjust the blank-holder force locally.
4. Regulating the type, amount, and distribution of the lubricant during sheet forming.

Hydroforming

Hydroforming is characterized by the replacement of a rigid tool through a pressure medium such as water or oil. There are two broad groups of these processes: Sheet hydroforming and tube hydroforming. Sheet hydroforming processes are basically of two types: processes in which the die is replaced by a pressurized fluid (called hydromechanical deep drawing, the hydromech process, the aquadraw process, hydraulic counter-pressure deep drawing, or just hydroform) and processes in which the punch is replaced by a pressurized fluid (called high-pressure sheet forming or just fluid forming).

High-Pressure Sheet Forming (Fluid Forming). The high pressure sheet forming process consists of two stages (Fig. 7.114): The free bulging stage and the cavity filling stage. Compared to conventionally deep-drawn parts, hydroformed parts have better tolerances and repeatability. Besides, parts made from blanks of different thickness and materials (tailored blanks) can be produced using the same tools. Another advantage is the low springback and low residual stress levels in the product. On the other hand, due to the free bulging the depths of the parts are limited. Also the cycle time is relatively slow, and dies cannot be changed as quickly as conventional dies.

For the given typical process in Fig. 7.114, the internal pressure at the end of the free bulging state can be estimated by the membrane theory as

$$p \approx 4\sigma_{f0} t_0 \frac{r_d}{a^2 + r_d^2} . \qquad (7.115)$$

The maximum height for the free bulging stage is given by

$$h_{max} = \frac{a}{\sqrt{5}} \approx 0.447a . \qquad (7.116)$$

The pressure to fill the cavities is approximately

$$p = \sigma_{f0} t_0 \left(\frac{1}{a} + \frac{1}{r_d} \right) . \qquad (7.117)$$

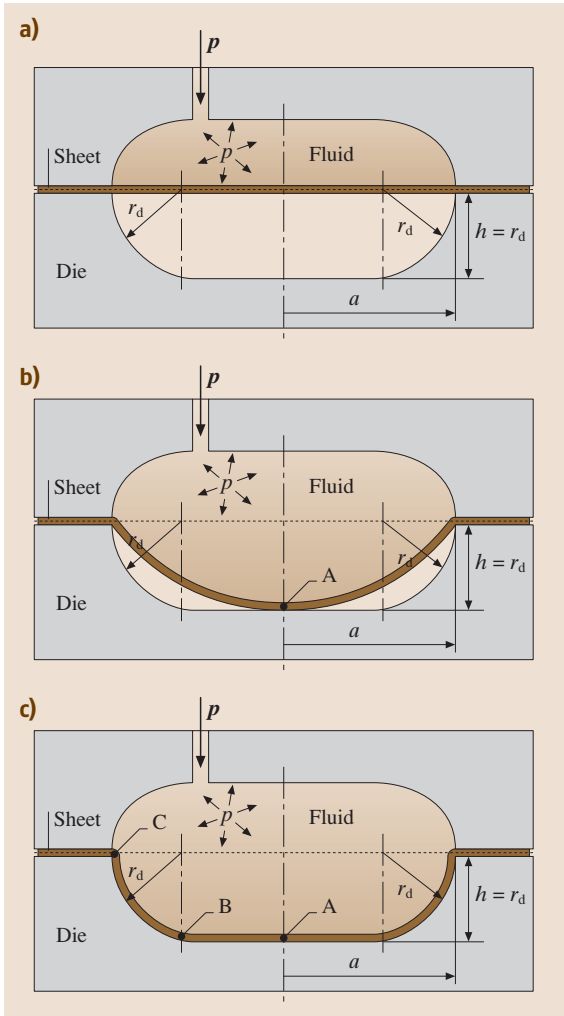


Fig. 7.114a–c Principle of high pressure sheet forming: **(a)** Beginning of process, **(b)** end of free bulging, **(c)** end of cavity filling (after [7.52])

The contact pressure in the interface of the sheet and the rigid die between point B and C is estimated by

$$p_{contact} \approx \frac{\sigma_{f0} t_0}{a} . \qquad (7.118)$$

Hydromechanical Deep Drawing. The working principle of hydromechanical deep drawing is given in Fig. 7.115 [7.53]. First, the press is opened and the water container is filled in the home position. After the insertion of the blank, the press closes, and the blank holder grips the blank. The blank holder pressure, set

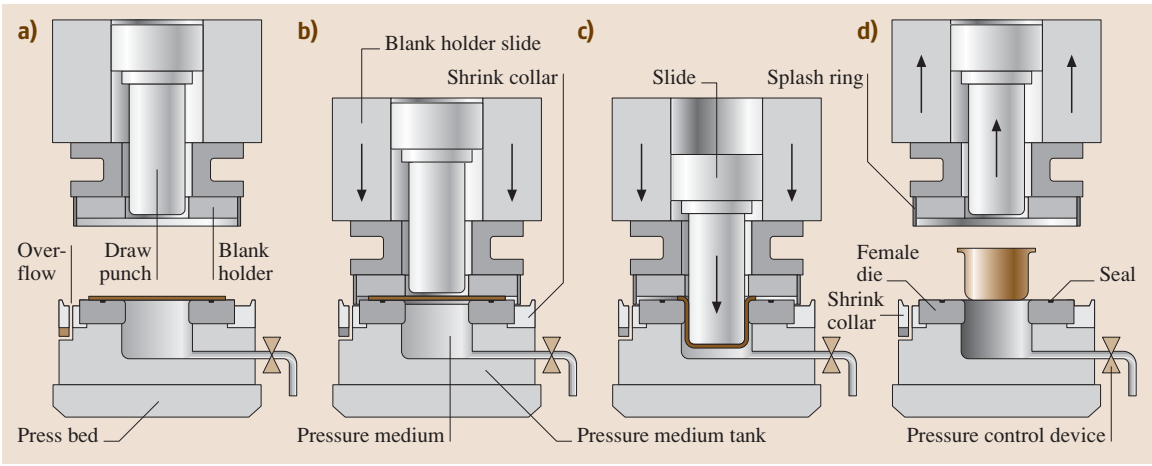


Fig. 7.115a–d Principle of hydromechanical deep drawing (after [7.53])

at the press, seals the pressure chamber, and the actual forming process is initiated. The medium pressure builds up as a result of penetration of the draw punch into the water container. During deformation, the sheet metal is pressed against the draw punch. During the forming phase, the control system, which is linked to the pressure chamber, controls the application of the hydraulic pressure in function of the draw depth.

After reaching the mechanically limited draw depth, the pressure in the chamber is released and the press travels back to its home position.

Advantages of hydromechanical deep drawing are the higher limiting drawing ratio ($\beta_{max} = 2.7$) as compared to mechanical deep drawing ($\beta_{max} = 2.0$), the better surface quality, and the lower springback of the product. Pressures for various materials during hydromechanical deep drawing are given in Table 7.32.

Tube Hydroforming. Tube hydroforming is a cold forming process in which a tube is pressurized internally and simultaneously compressed axially and/or radially [7.53]. The tube is thereby expanded and pressed against a die. The process steps can be illustrated by the manufacturing of a T-fitting (Fig. 7.116): A special hydraulic press is equipped with a two-part multiple purpose die. Depending on the workpiece, the dies have two seal punches (horizontal cylinders) positioned axially relative to the tube ends and a counterpressure punch. The tubular preform is in the bottom die and the die is closed. The ends of the tube are sealed by the axial punches, and the tube is filled with pressure medium. In the actual forming process, the punches compress the tube, while the pressure medium is fed to inflate the part until the part wall rests against the die contour. Pressures can reach values up to 4000 bar. The counterpressure punch additionally controls the material flow. The calibration pressure forms the workpiece in such a way that its contour corresponds to that of the die accurately and reproducibly. The die is finally opened and the formed component is ejected.

Currently, tubes with maximum dimensions given as in Fig. 7.117 are possible to form by tube hydroforming. Typical achievable product dimensions are given in Fig. 7.118.

Tube hydroforming is limited by three failure modes (Fig. 7.119): buckling, wrinkling, and bursting.

Table 7.32 Pressures during hydromechanical deep drawing

Material	Pressure range (bar)
Aluminum	50–200
Steel	200–600
Stainless steel	300–1000

7.2.6 Forming Machines

Metals are plastically formed by usually at least two tools. These tools have to be moved with certain kinematics and with certain forces. This is achieved by a forming machine. Since various metal forming processes need various load-stroke characteristics, the forming machines used have to be selected accordingly. Besides satisfying the mechanical requirements, forming machines must also provide a long uptime, i. e. the time they are used for forming, and they should ensure

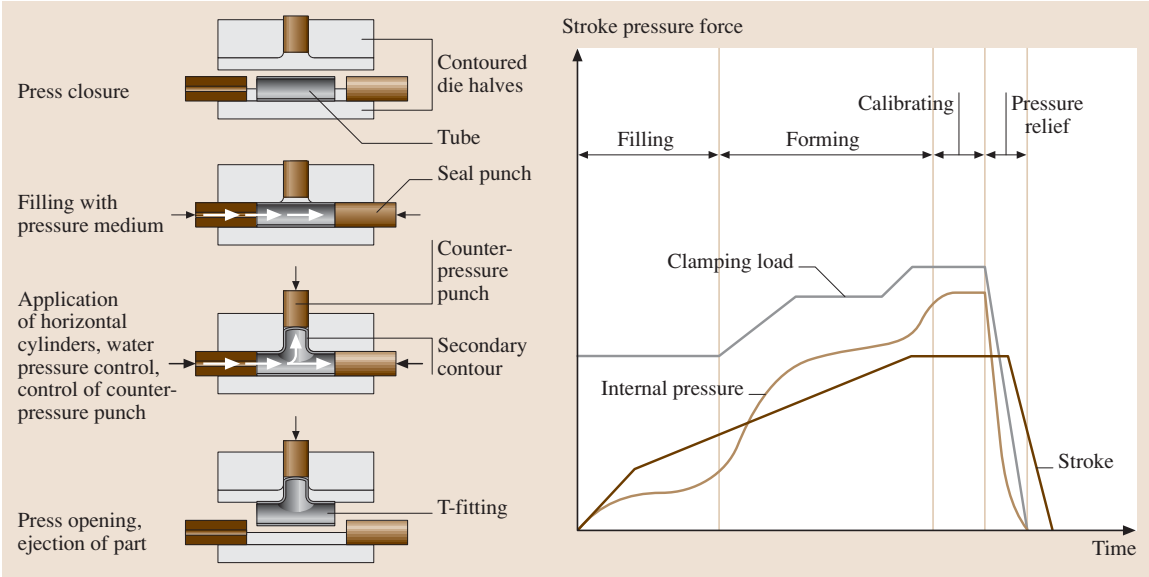


Fig. 7.116 Principle of tube hydroforming and process control (after [7.53])

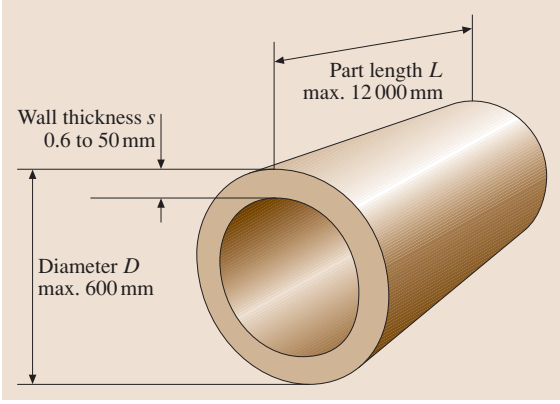


Fig. 7.117 Current limits on workpiece geometry for tube hydroforming (after [7.53])

long tool lives. The latter depends, for instance, on the precision of the guidance of the slides.

The largest group of forming machines is presses. These are machines with linear relative motion between the two tools. Other forming machines include rolling machines, wire or tube drawing machines, and magnetic or explosive forming machines. Here only presses will be discussed.

Parameters of Forming Presses

The various requirements of the forming process such as necessary force, moment, energy, etc. must be matched with the capabilities of the forming press. This can be achieved by comparing characteristic parameters of the forming press with those of the forming process. Basically four groups of parameters can be identified [7.22, 23]: energy and force parameters, transient parameters, accuracy parameters, and other parameters.

Energy and Force Parameters. For a feasible forming process the nominal ram force provided by the machine F_{nom}, must be larger than or equal to the necessary maximum forming force F_{form} and the available energy of the forming machine E_{nom} must be larger than or equal to the necessary process work W_{proc}. Practically, however, equality makes no sense due to the inherent fluctuations of the forming loads; on the other hand, due to economic utilization of the usually expensive forming machines F_{nom} and E_{nom} should be as close as possible to F_{form} and W_{proc}, respectively. The necessary process work is given by

$$W_{proc} = W_{form} + W_{el} + W_{fr} + W_{aux} \,, \qquad (7.119)$$

where W_{form} is the work necessary to plastically deform the metal (including the frictional work at the tool/workpiece interface), W_{el} is the work to elastically deform the press frame and the tools, W_{fr} is frictional work consumed in the various guides and bearings, and finally W_{aux} is the work consumed by auxiliary equipment such as blankholders, ejectors etc.

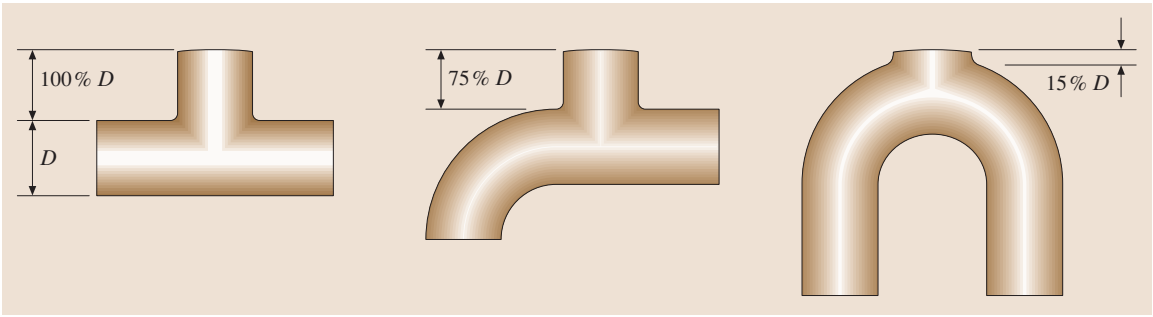


Fig. 7.118 Producible product geometries by tube hydroforming (after [7.53])

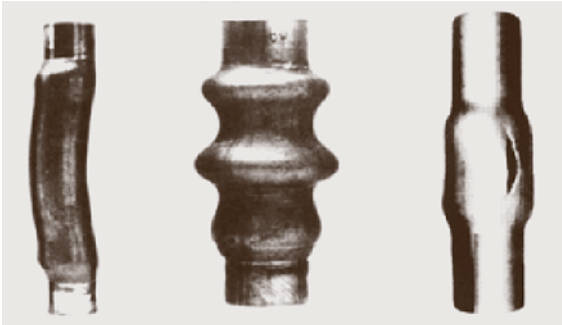


Fig. 7.119 Failure modes in tube hydroforming (after [7.54])

The elastic work absorbed by the press frame and the tools is given by

$$W_{el} = \frac{1}{2}\frac{F_{form}^2}{C}\,, \qquad (7.120)$$

where C is the stiffness of the forming press including the tools. As long as $C > dF_{form}/ds$ (with s the ram stroke), the elastic energy spent until the maximum forming load is reached can be recovered afterwards. The not recovered elastic energy leads to vibrations in the drives and in the machine frame. Hence, the higher C is, the higher the chance that the elastic energy can be recovered and vibrations can be reduced.

Presses can be classified according to the basic physical parameter controlling the action of the press as stroke controlled, force controlled, and energy controlled presses (Fig. 7.120): *Energy controlled forming machines* provide a certain amount of energy (E_{nom}) for the process. If this energy is consumed the press stops. If the required energy for the process is larger than the available one multiple strokes can be executed (Fig. 7.120a). Hence the basic characteristic parame-

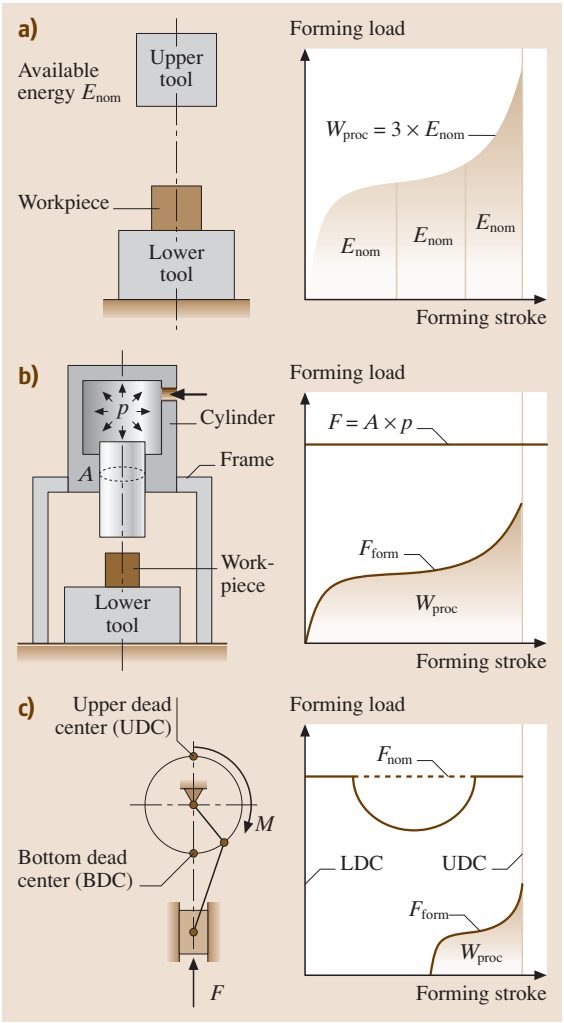


Fig. 7.120a–c Basic types of forming presses: **(a)** Energy controlled forming presses, **(b)** load controlled forming presses, **(c)** stroke controlled forming presses

ter of energy controlled presses is the nominal energy E_{nom}. Hammers and screw presses are two typical representatives of this group. Different from hammers, screw presses have drives and frame elements under load, so that for these in addition to the nominal energy also a nominal force has to be specified for which these machine elements are designed. *Force controlled presses* provide independent of the stroke a given force that is the obtained by the hydraulic pressure p multiplied by the cylinder cross-sectional area A (Fig. 7.120b). The basic parameter for these presses is therefore the maximum allowable force F_{nom}. The typical representative of this group is the hydraulic press. Finally, *stroke controlled presses* (Fig. 7.120c) provide ram force for each ram position depending on the kinematics of the mechanical drive. The characteristic parameters, therefore, are the ram force as a function of the stroke and the maximum ram force F_{nom}. Typical representatives are crank and toggle presses.

Transient Parameters. The basic time-dependent parameters are the effective stroke rate, the contact time of the tools with the workpiece under the forming load, and finally the speed of the press. The effective stroke rate determines the economic efficiency of the press. This parameter is related to the failing height in case of hammers, to the speed during the load free stroke in case of hydraulic presses and to the speed and total stroke in case of crank presses. The contact time under pressure is especially important for warm and hot forming processes since it determines the cooling amount of the workpiece. Typical contact times for various machines are given in Table 7.33 For stroke controlled presses the contact time under pressure is larger the softer the frame and tool system are (i. e. the lower C is). Another important parameter is the speed of the ram of the press. This directly influences the strain rates during forming. Especially in warm and hot forming, the higher the strain rates, the higher the flow stress, and hence the higher the forming loads. In hammers the speed is given during the forming process by the power balance, whereas for stroke driven presses the ram speed is a function of the ram position. For the latter, it must be noted that the true ram speed depends on the stiffness of the frame-tool system.

Accuracy Parameters. The accuracy parameters of presses are related to geometric errors of the workpiece, such as position errors during impact, eccentricity of the product, dimensional errors in product height, the angle of twist of the product, etc. These errors are caused either by inaccuracies of the presses in the idle state, such as excessive clearance of the guides or skewness of the lower die and upper die leading to position errors during impact, or by inaccuracies of the press under load, such as elastic deformations leading to height errors in the tool. The latter errors are strongly dependent on the stiffness of the press, so that this characteristic parameter is the key parameter for the press specification. In presses generally the frame, the upper tool, the lower tool, and the drive system deflect elastically. In the case of hammers, only the lower tool deflects elastically.

Table 7.33 Order of magnitude for contact times for various press types (after [7.22, 23])

Press type	Contact times under pressure (s)
Hammers	10^{-3}–10^{-2}
Screw presses (with fly-wheel)	10^{-2}–10^{-1}
Stroke controlled presses	10^{-1}–5×10^{-1}
Hydraulic presses	10^{-1}–1

Other Parameters. Finally, parameters such as stroke length, tooling space, space requirements of the press, weight of the press, and necessary power supply can be listed as other characteristic parameters.

Energy Controlled Presses

Energy controlled presses provide a certain predetermined amount of energy for the forming process. Force as well as displacement is not controlled directly. There are basically two types of energy controlled presses: Hammers and screw presses.

Hammers. There are basically three types of hammers (Fig. 7.121): Drop hammers provide the energy by a freely falling ram including the upper die. In double-acting hammers the ram is accelerated by a fluid such as steam, air, or hydraulic oil acting through a cylinder and piston. Finally, in counter-blow hammers upper and lower dies are accelerated towards each other. The first two types transmit the forging force to the ground, whereas for the counter-blow hammer the ground is practically not effected by the forming load.

The properties of these hammers are given in Table 7.34. Hammers are basically used in hot forging operations, so that the impact speeds are important for determining the strain rates for the forming process.

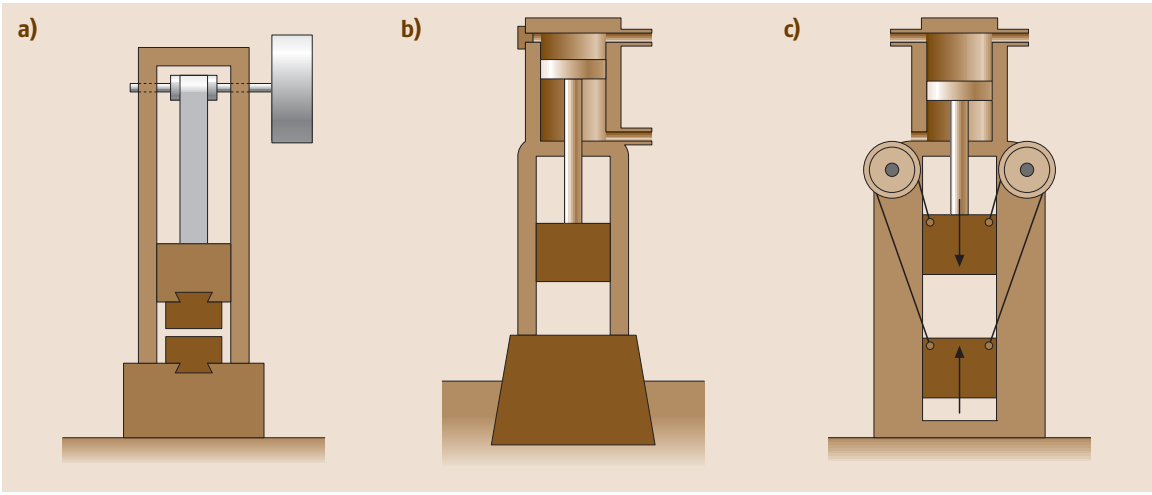


Fig. 7.121a–c Basic types of hammers: (a) Drop hammer, (b) double-acting hammer, (c) counter-blow hammer

Furthermore, the contact time of the dies with the workpiece is important to determine the cooling of the workpiece. The contact time for hammers is 10^{-3} to 10^{-2} s and is the shortest contact time among all press types. The maximum nominal energy of the hammer cannot be completely used for forming due to energy losses such as friction at the slides and elastic deformations of the dies. For small loads and large displacements about 80 to 90% of the nominal energy can be used, whereas for high loads and small displacements only 20 to 50% of the nominal blow energy is available for the forming process.

Screw Presses. Screw presses have two basic types (Fig. 7.122): presses with axially moving screw drives and presses with stationary screw drives (also called Vincent presses). The angular kinetic energy of the flywheel is transferred to the screw that either moves the upper ram (Fig. 7.122a) or the lower die (Fig. 7.122b). In the case of Vincent presses, the frame of the press is not under load during forming. The available energy is given by the flywheel rotational speed ω (in rad/s) and its moment of inertia I as

$$E = \frac{1}{2} I \omega^2 \,. \tag{7.121}$$

This energy is used for forming, for overcoming friction in the slides, and for elastically deforming the frame of the press.

The contact time of the dies with the workpiece is longer for screw presses than for hammers and is between 10^{-2} to 10^{-1} s. The impact speed of the dies on the workpiece ranges between 0.7 to 1 m/s. The maximum nominal energy ranges from 800 to 7500 kN m and the maximum nominal press loads from 20 000 to 140 000 kN, depending on the drive to accelerate the

Table 7.34 Properties of various hammers (after [7.38])

Property	Drop hammers	Double-acting hammers	Counterblow hammers
Maximum drop height	1.3–2 m	about 1.3 m	–
Acceleration of ram	$< g$	$> g$	$> g$
Impact speed of ram	3–5 m/s	3–9 m/s	5–14 m/s
Maximum nominal energy E_N	80–160 kN m	200 kN m	1000 kN m
Blow energy equation	mgh	$mgh + pAh$	$\frac{1}{2}(m_1 + m_2)v^2$
Parameters	m: mass of ram, g: gravitational acceleration, h: drop height, p: pressure in cylinder, A: cross-sectional area of cylinder, m_1: mass of upper ram, m_2: mass of lower ram, v: speed of each ram		

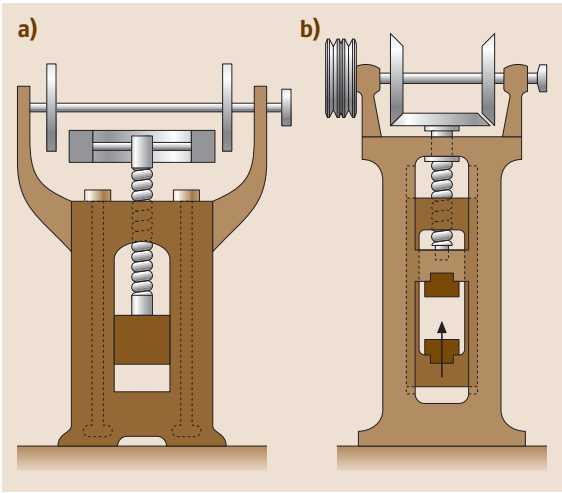


Fig. 7.122a,b Basic types of screw presses: **(a)** Friction screw press, **(b)** Vincent friction screw press

flywheel (friction, gear, electrical, or hydraulic drives). Impact loads are between 40 000 to 300 000 kN.

Compared to hammers screw presses have a lower noise level. Because of the lower hitting speed the strain rates are lower, yet the cooling time of the workpiece under pressure is longer. Also for the same energy, screw presses have a flywheel mass that is about 1/6th of the mass of the upper die of a hammer. Furthermore, the accuracy of screw presses is higher than the accuracy of hammers. A disadvantage of screw presses is, however, that their stroke rate is lower than that of hammers. Screw presses are, therefore, also used in cold calibration operations and coining in addition to hot forging.

Stroke Controlled Presses

Stroke controlled presses are also simply called mechanical presses. They all utilize a flywheel to store the necessary forming energy. Figure 7.123 shows a typical eccentric press. A slider crank mechanism is used to transfer the energy of the flywheel through the crank, the connecting rod, and finally the slide to the workpiece. A slight variation of the eccentric press is the crank press, where a crank link is used instead of an eccentric shaft. The presses are sometimes differentiated by the fact that the total stroke of the crank press cannot be changed whereas the total stroke of the eccentric press is adjustable [7.42].

For crank presses the ratio of the crank length to coupler length r/ℓ takes values between 1/15 to 1/4, so

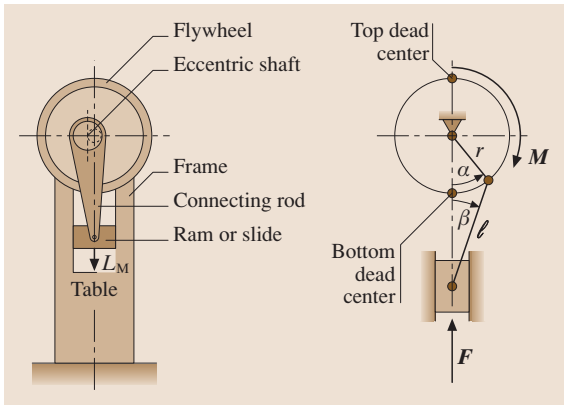


Fig. 7.123 A typical eccentric press and its mechanism

that the included angle between the coupler link and the press axis β is less than 6°. For this case, the axial ram force acting on the slider is given by

$$F(\alpha) \approx \frac{M}{r \sin \alpha} , \tag{7.122}$$

where M is the torque supplied by the crank and r is the crank radius. The nominal force of a mechanical press is given for a crank position at 30° to its bottom dead center

$$F_{\text{nom}} = F(\alpha = 30°) . \tag{7.123}$$

This is the highest allowable forming force for the given press. All the mechanical elements of the press are designed for this load. For various crank angles the current allowable force and the theoretical available force is shown in Fig. 7.124. The crank angle dependence can be converted to the stroke of the press s as measured from the bottom dead center using

$$s = r \left[(1 - \cos \alpha) + \frac{1}{2} \left(\frac{r}{\ell} \right) \sin^2 \alpha \right] . \tag{7.124}$$

If the total stroke of the press is denoted by S $(= 2r)$, then the stroke for which the nominal force is available corresponds to 7.3% of S for $r/\ell = 0.1$. Beyond this rather small stroke the available ram force decreases rapidly.

A another force-stroke characteristic is given for knuckle-joint (toggle) presses (Fig. 7.125). The six-link mechanism employed in this press provides the nominal force over a much smaller stroke than the crank presses, so that they are suitable basically for processes with force maximums toward the end of the stroke re-

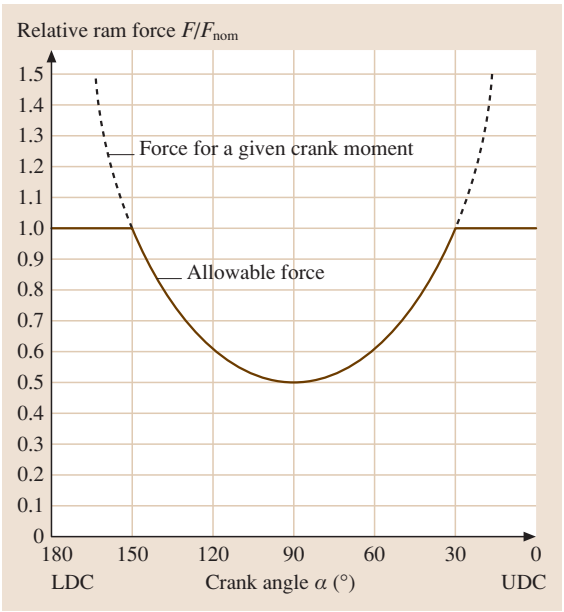


Fig. 7.124 Allowable press force as a function of crank position

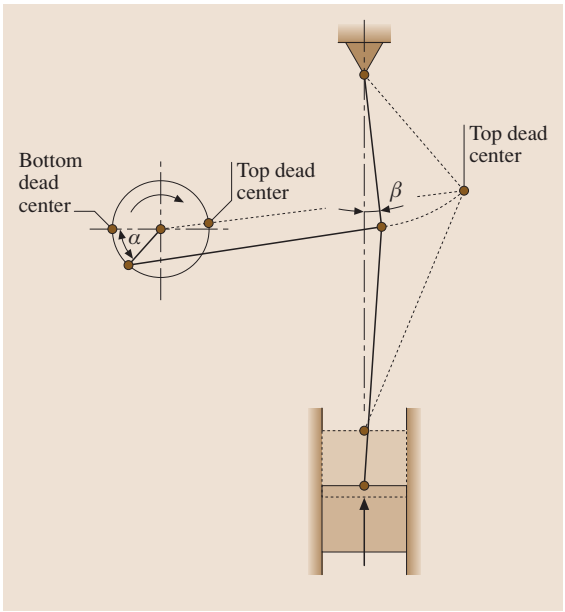


Fig. 7.125 Mechanism of a knuckle-joint press

quiring also short strokes. Despite this disadvantage, the velocity characteristic of toggle presses is superior to crank presses since they are faster over a long stroke but slower towards the dead centers, i. e. in the forming regime, than crank presses (Table 7.35).

A recent development in press technology is the servomotor press [7.42]. One or multiple servomotors are used to drive a crank or toggle mechanism. The obvious advantage of the servomotor press is the almost limitless control of the press. Current limitations are the available motor torques and the hence forming forces, since the forming energy must be supplied instantaneously. Despite this disadvantage 1000 t presses are expected to be available in near future.

Force Controlled Presses

Hydraulic presses supply the nominal forming force independently from the stroke. Figure 7.126 shows the basic working principle of a hydraulic press: A pump converts the electrical energy into hydraulic energy that is converted finally to mechanical energy. A cylinder in the die closure phase is pressurized by hydraulic fluid from a reservoir with high flow rates. This fluid is usually oil. This oil is compressible such that a 0.7 to 0.8% volume change occurs per 100 bar pressure. After the dies are closed control of the cylinder is overtaken by the slide control for the forming action. The cylinder is returned by reversing the fluid flow into the lowering cylinder. Hydraulic presses have the advantage that the whole cycle of die-closure/forming/punch reversal can be programmed efficiently and individually. Therefore, these press types are appropriate for many forming processes, such as sheet forming, cold forging, and especially for hot forging and hot extrusion.

The basic characteristic parameter of a hydraulic press is the nominal forming force that is obtained by

Table 7.35 Properties of mechanical presses (after [7.38])

Property	Eccentric or crank press	Knuckle-joint press
Stroke	10–300 mm	3–12 mm
Nominal load	1000–16 000 kN	1000–16 000 kN
Stroke rate	10–100 strokes/min	20–200 strokes/min

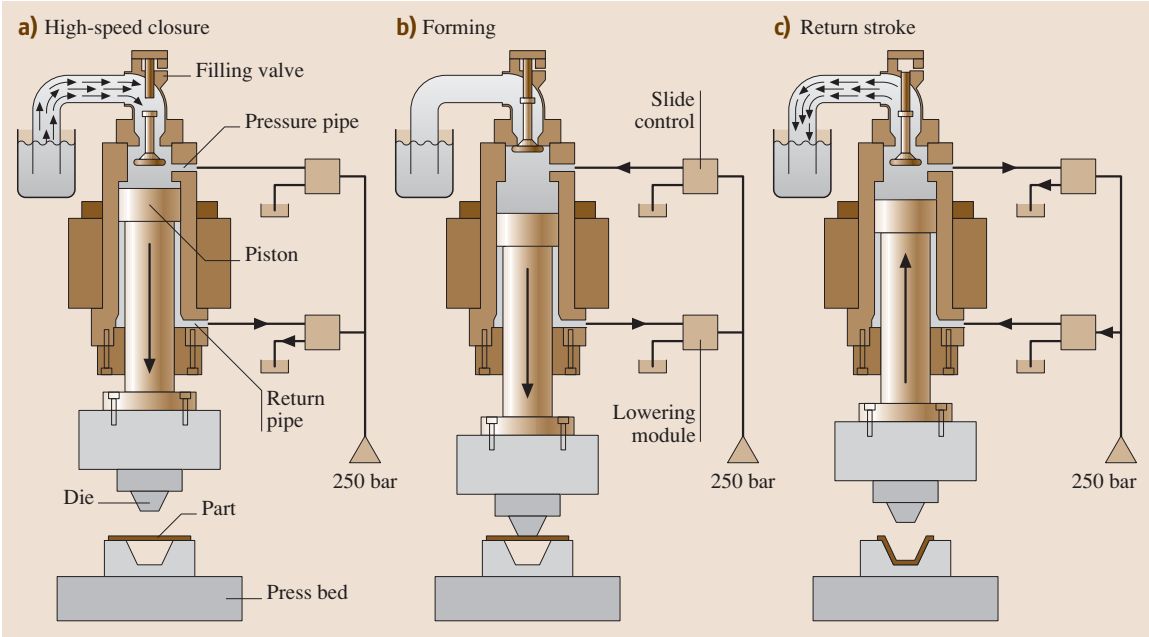


Fig. 7.126a–c Working principle of a hydraulic press: **(a)** High-speed closure, **(b)** forming, **(c)** return stroke (after [7.53])

the hydraulic pressure p and the cross-sectional area of the cylinder A

$$F_{\text{nom}} = pA \,. \tag{7.125}$$

The hydrostatic pressure is up to 320 bar. In contrast to mechanical presses there is no energy reservoir, so that the necessary energy during forming must be supplied instantaneously by the pump. This pump power P is given by

$$P = \frac{1}{\eta_{\text{total}}} p\dot{V} \,, \tag{7.126}$$

where $\dot{V}$ is the volumetric flow rate of the fluid and η_{total} is the total efficiency of the hydraulic system including mechanical and electrical losses. The maximum flow rates are around 185 l/min.

7.3 Machining Processes

In material removal processes surplus material is removed from a solid object (workpiece) in the form of small pieces (e.g. chips) by means of tools, thus generating surfaces by relative motions between the workpiece and tool provided by the machine tool.

The material removal processes that will be discussed in this section can be divided into three subgroups:

- Cutting with geometrically well-defined tool edges
- Cutting with geometrically undefined tool edges
- Nonconventional machining (chipless) processes

Cutting is accomplished with the mechanical action of a tool on a workpiece. In chipless machining material particles are removed from a solid object by nonmechanical means.

7.3.1 Cutting

Fundamentals

During a cutting process particles of material are mechanically removed as chips from a blank or workpiece by the cutting action of a tool. In machining with geometrically well-defined tool edges the number of cutting

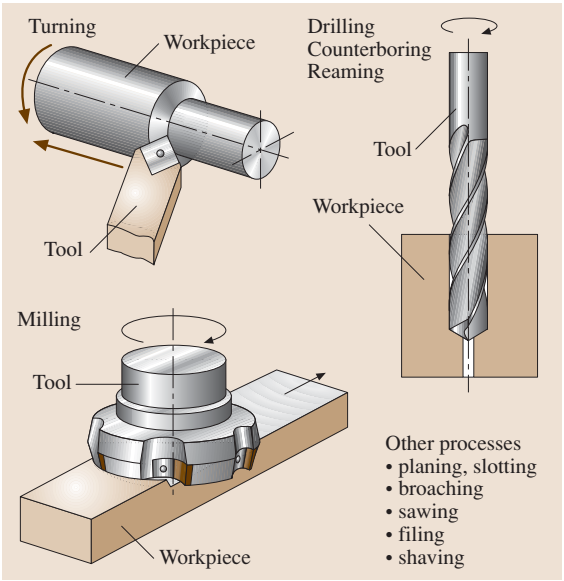


Fig. 7.127 Cutting processes (machining with geometrically well-defined tool edges) [7.55]

edges, the geometry of the tool and the position of the cutting edges in relation to the workpiece are known and describable (as opposed to cutting with geometrically undefined tool edges, e.g. grinding).

Figure 7.127 [7.55] illustrates important processes of this group. The processes differ according to the cutting motion (cutting speed v_c), the feed motion (feed rate v_f), and the resulting effective motion (effective cutting speed v_e). Feed and cutting direction vectors span the working plane. The angle between the two vectors is called the feed motion angle φ, while the angle between the effective direction and the cutting direction is called the effective cutting speed angle η. The following equation applies for all processes

$$\tan \eta = \frac{\sin \varphi}{\frac{v_c}{v_f} + \cos \varphi} \,. \qquad (7.127)$$

The mechanical process of separation of material parts from the workpiece, i. e. chip formation, can be described by the example of an orthogonal process (two-dimensional deformation). The wedge is described by the rake angle γ, the clearance angle α, and the edge radius r_B. The wedge penetrates into the working material, which is plastically deformed. Figure 7.128 [7.56] shows the zones of plastic deformation in continuous chip formation. Five effective zones can be distinguished.

The primary shear zone comprises the actual area of chip formation by shearing. In the secondary shear zones in front of the rake and flank face, friction forces act between the tool and the workpiece causing plastic deformation of these material layers. In the zone in front of the deformation area, chip formation results in stresses that lead to plastic and elastic deformations. In the pressure and cutting zone, the material is deformed and separated under high compressive stresses.

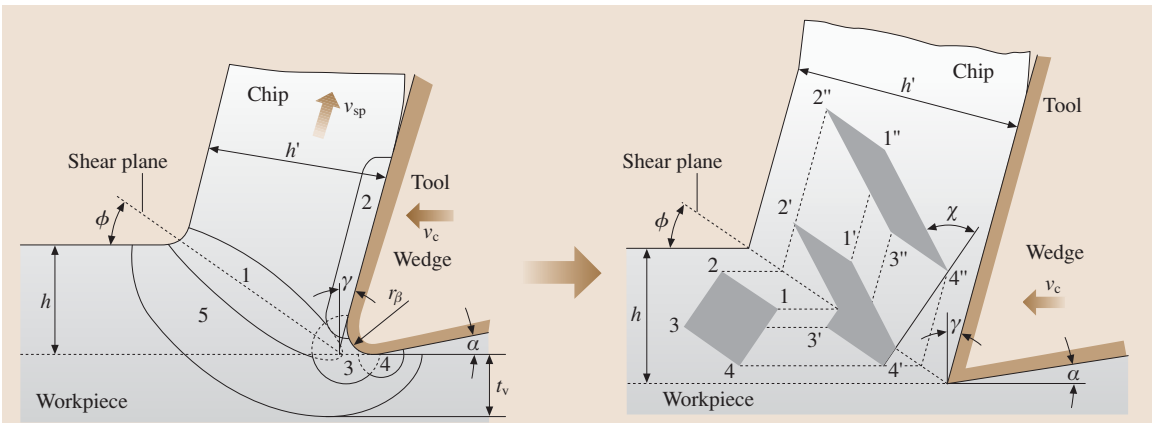


Fig. 7.128 Effective zones in chip formation and derived model of deformation in the shear plane

In the course of these processes the undeformed chip thickness h gives way to the chip thickness h' with the resultant chip compression ratio being $\lambda_h = h'/h$. The shear plane comprises the shear angle ϕ with the cutting speed vector. The deformation angle χ characterizes the shearing of a small part that has passed through the shear plane.

The following chip types can be distinguished: in continuous chip formation the chip slides off over the rake face at a constant speed in a stationary flow. Periodical changes in the intensity of the deformation may occur; mainly at higher cutting speeds. Lamellae are formed in the chip that can reach as far as material separation and the formation of chip segments.

Segmented chip formation is the discontinuous formation of a still continuous chip that, however, shows recognizable differences in the degree of deformation along the direction of flow. Especially negative tool orthogonal rake angles, but also greater undeformed chip thicknesses, as well as very slow and very fast cutting speeds can result in segmented chip formation. The chip formation is uneven.

Discontinuous chip formation occurs in working materials that possess only a small ability to deform, such as cast iron with graphite lamellae. The parting surface between the chip and the workpiece proceeds irregularly.

Built-up edges can form in ductile, work-hardening materials at low cutting speeds and sufficiently continuous chip formation. Parts of the working material, that have been strongly deformed and cold-hardened in the area of the compression zone weld under high pressure at the rounded cutting edge to the chip surface, thus becoming a part of the cutting tip.

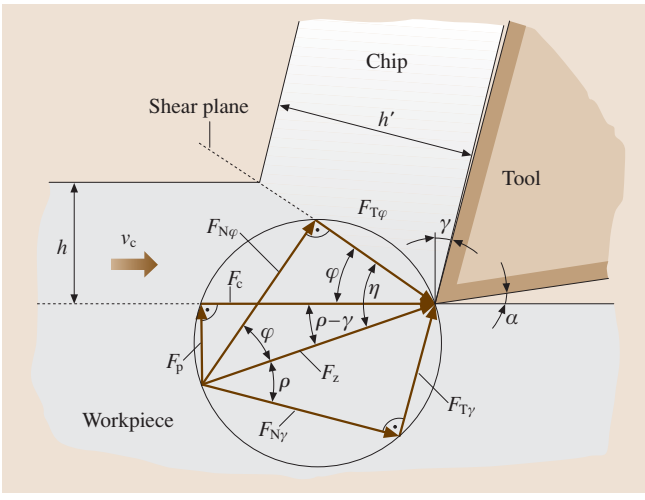


Fig. 7.129 Ernst and Merchant's force circle

The forces acting on the tool in orthogonal cutting can be represented by Ernst and Merchant's force circle as shown in Fig. 7.129 [7.57]. The resultant force F_z has two components, F_c and F_p. The cutting force F_c in the direction of the tool path determines the amount of work done in cutting. The thrust force F_p, together with F_c, produces deflections of the tool. The resultant force has two components on the shear plane: $F_{T\varphi}$ is the force required to shear the metal along the shear plane, and $F_{n\varphi}$ is the normal force on this plane. Two other force components on the face of the tool are also represented: the friction force $F_{T\gamma}$ and the normal force $F_{N\gamma}$. From the geometries within the circle the following relationships can be derived.

The frictional force on the tool rake face

$$F_{T\gamma} = F_p \cos\gamma + F_c \sin\gamma \,. \tag{7.128}$$

The coefficient of friction at the tool-chip interface

$$\mu \quad = \frac{F_p + F_c \tan\gamma}{F_c - F_p \tan\gamma} \,. \tag{7.129}$$

The shear stress in the shear plane, where $A_0 = hw$ is the undeformed chip cross-section and w is the width of the cut

$$\tau = \frac{F_c \sin\varphi \cos\varphi - F_p \sin^2\varphi}{A_0} \,. \tag{7.130}$$

In the chip formation zone the cutting energy applied E_c is completely converted. It can be calculated from

$$E_c = F_c l_c \,, \tag{7.131}$$

where F_c is the cutting force and l_c is the travel length in cutting direction.

The cutting energy is composed of deformation and shear energy E_φ, friction energy at the rake face E_γ, friction energy at the flank face E_α, surface energy for the formation of new surfaces E_τ and kinetic energy due to chip deflection E_M.

The energy converted during machining one unit of volume is

$$e_c = \frac{E_c}{V_w} \,, \tag{7.132}$$

where e_c is the specific energy and V_w is the volume of material removed.

Like E_c, itself, its individual components can be expressed in relation to V_w.

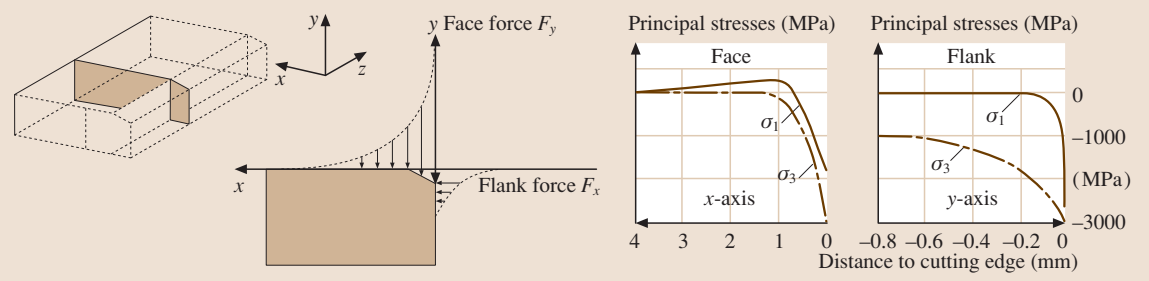


Fig. 7.130 Load distribution due to mechanical stress perpendicular to the major cutting edge

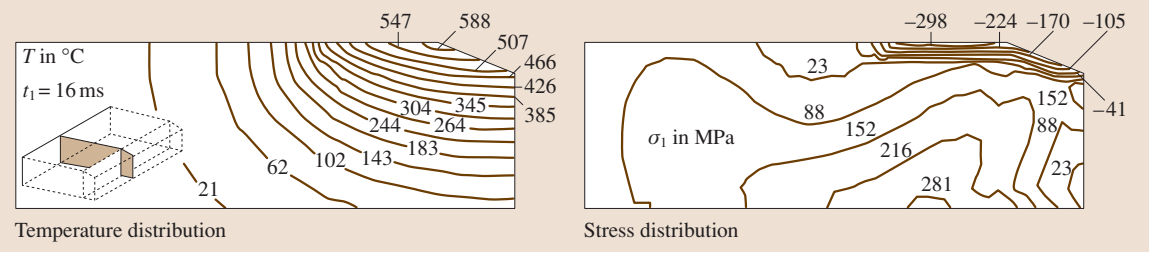


Fig. 7.131 Temperature and stress distribution during constant heat flow to the ceramic cutting tool

From the specific energy e_c the specific cutting force k_c can be derived as a characteristic value for calculating the cutting force

$$k_c = \frac{F_c}{A} = \frac{F_c}{hb}\,, \qquad (7.133)$$

where A is the undeformed chip cross-section, b is the undeformed chip width, and h is the undeformed chip thickness

$$e_c = \frac{E_c}{V_w} = \frac{P_c}{Q_w} = \frac{F_c v_c}{A v_c} = k_c, \qquad (7.134)$$

where P_c is the cutting power, Q_w is the material removal rate and F_c is the cutting force.

Estimation shows that the largest amount of cutting energy is converted into deformation and frictional energy. Thus the specific cutting force k_c can be seen as an energy-related variable (the application and determination of k_c will be dealt with in more detail in the section *Turning*). The energy introduced into the chip formation zone is almost entirely converted into heat, the minor amount being transformed into residual stress in the chip and in the workpiece (spring energy). Due to this process high temperatures in the cutting edge arise, introducing mechanical and thermal load into this area. Surface forces beneath the rake and flank face surfaces, and the principal stresses that can be calculated thereby are shown in Fig. 7.130 [7.56].

Figure 7.131 [7.56] represents the temperature distribution on a highly loaded ceramic indexable insert and the resulting thermally induced tensile stresses. These tensile stresses are particularly critical for high-temperature ceramic cutting materials. Mechanical and thermal loads supported by chemical reactions may cause significant wear.

The characteristics of worn cutting tools can be described by different types of wear (Fig. 7.132) [7.56]:

- Fractures and cracks. These arise in the area of the cutting edge because of mechanical or thermal overload.
- Mechanical abrasion (frictional wear) is mainly caused by hard inclusions such as carbides and oxides in the working material.

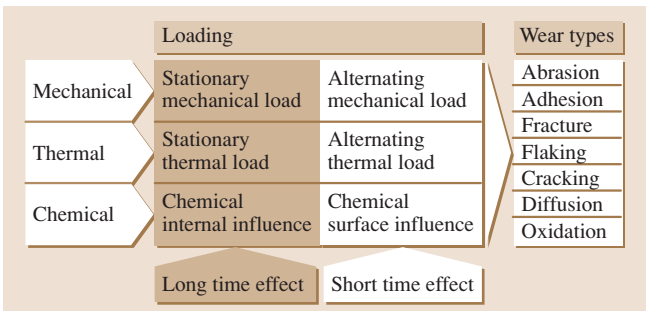


Fig. 7.132 Loading and wear types of cutting tools

- Plastic deformation occurs if the cutting material offers inadequate resistance to deformation but provides sufficient toughness.
- Adhesion is the shearing off of pressure welding areas between the working material and the chip, whereby the point of shearing is located in the cutting material.
- Diffusion occurs at high cutting speeds and mutual solubility of the cutting material and the working material. The cutting material is weakened by chemical reactions, dissolves and is removed.
- Oxidation also occurs only at high cutting speeds. By contact with oxygen in the air the cutting material oxidizes and the structure is weakened.

The machinability of a workpiece is determined by the composition of the workpiece material, its structural configuration in the machined area, its previous casting or forging process, and its heat treatment. Machinability is evaluated considering the following criteria:

- Tool wear
- Surface quality of the workpiece
- Machining forces
- Chip form

For the evaluation of an individual criterion the machining task must be considered.

Turning

Turning is defined as machining with a continuous (usually circular) cutting motion and any desired feed motion in a plane perpendicular to the cutting direction. The turning axis of the cutting motion keeps its position relative to the workpiece independently of the feed motion. Figure 7.133 [7.58] shows some important turning processes.

In the following, longitudinal cylindrical turning is taken as an example of a turning process. Terminology, names, and designations for geometrical descriptions of the cutting processes can be found in ISO 3002/1. Figure 7.134 [7.56] shows the surfaces and cutting edges defined for the cutting tip.

The angles represented in Fig. 7.135 [7.59] serve to determine the position and shape of the tool in three dimensions: the tool cutting edge angle κ is the angle between the primary cutting edge plane and the working plane. The tool included or edge angle ε is the angle between the primary and secondary cutting edge planes and is given by the cutting edge geometry. The tool cutting edge inclination λ is the angle between the cutting edge and the reference plane and is apparent when looking down onto the primary cutting edge. The clearance angle α, wedge angle β, and rake angle γ can be measured in the tool orthogonal plane and make up 90° in sum. The values of the relevant tool angles are determined from approximate value tables in relation to

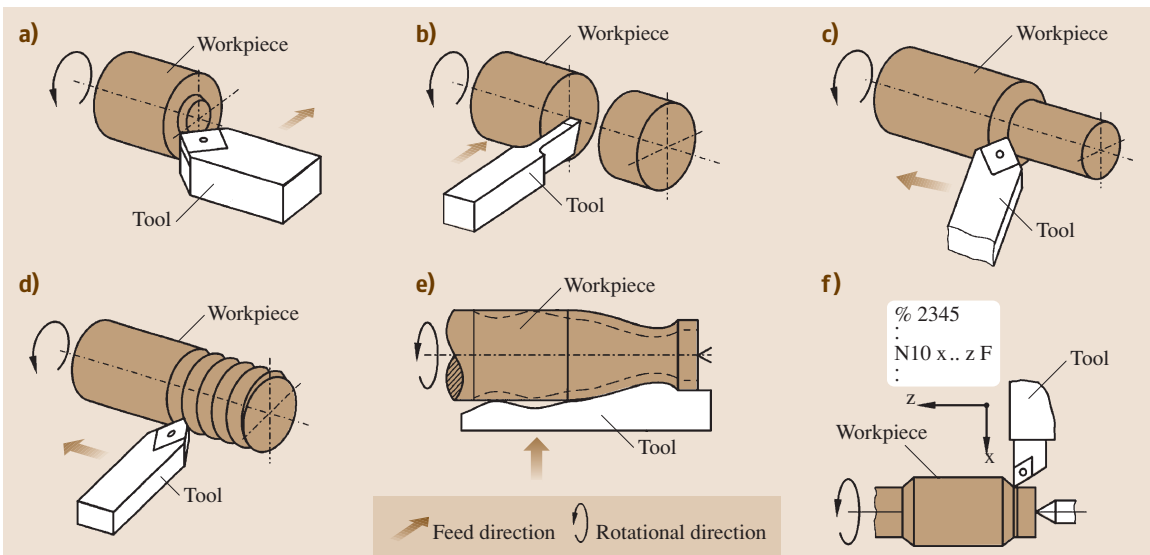


Fig. 7.133a–f Turning processes: **(a)** Facing, **(b)** parting-off, **(c)** longitudinal turning, **(d)** thread turning, **(e)** profile turning (tool contour is duplicated in workpiece), **(f)** form turning [7.58]

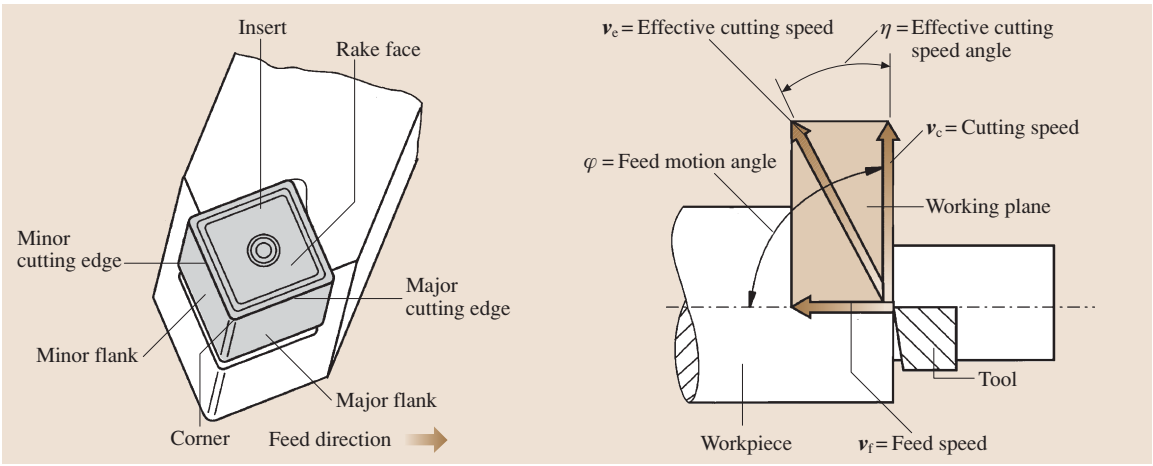


Fig. 7.134 Terminology at cutting tip and directions of motion of tool (ISO 3002/1)

the workpiece and cutting materials and the machining process. The tool cutting edge angle κ influences the shape of the undeformed chip cross-section to be removed and thus the power required for the machining process (Fig. 7.136) [7.56].

The bulk volume Q' of the chips flowing over the rake face of the tool varies depending on chip type and form. The characteristic parameter is the chip volume ratio RZ specifying the relationship between the removed chip volume in a given time Q' (bulk chip volume) and the material removal rate Q_w. In this case

$$RZ = \frac{Q'}{Q_w}\,, \tag{7.135}$$

$$Q_w = a_p F_c = a_p f D \pi n\,. \tag{7.136}$$

The chip volume ratio RZ characterizes the *bulkiness* of the chips. It serves to determine the machine tool working spaces, the chip conveying devices, and chip spaces in the cutting tools. The chip volume ratio can amount to very different values according to the chip form (Fig. 7.137) [7.60]. The more brittle the material, the lower this value. Brittleness can be influenced via the composition of the material. For steel, higher sulfur contents (more than 0.04%, free cutting steel with 0.2% S) have a beneficial effect. However, this may impair the toughness of the material in the transverse direction, depending on the form of the dispersed sulfides. Chip breakers ground into the chip surface, sintered in, or attached with the clamping system of inserts cause additional chip deformation, i. e. an additional material load in the chip. The chip is bent by contact with the cut surface of the workpiece or the flank face of the tool and breaks (secondary chip breaking in contrast to segmented or discontinuous chip formation, where the chips leave the chip formation zone as small fragments). Favorable chip forms can also be achieved by selecting the appropri-

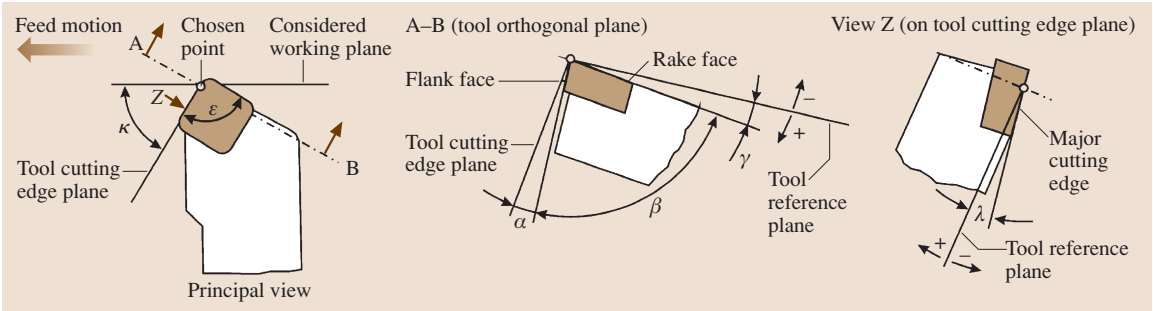


Fig. 7.135 Single point cutting tool angles (DIN 6581)

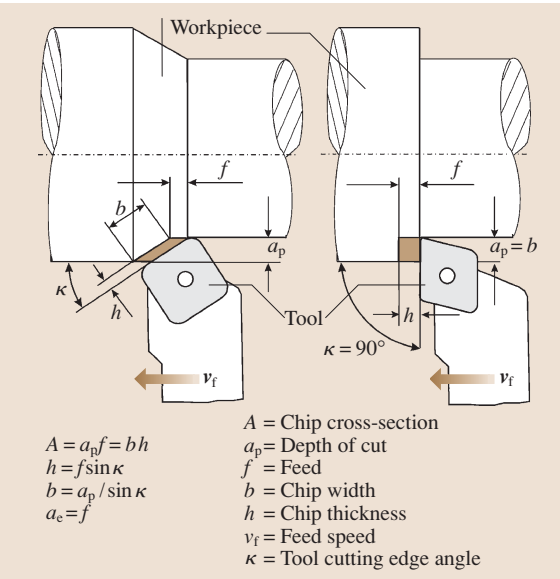


Fig. 7.136 Cut and chip variables in turning

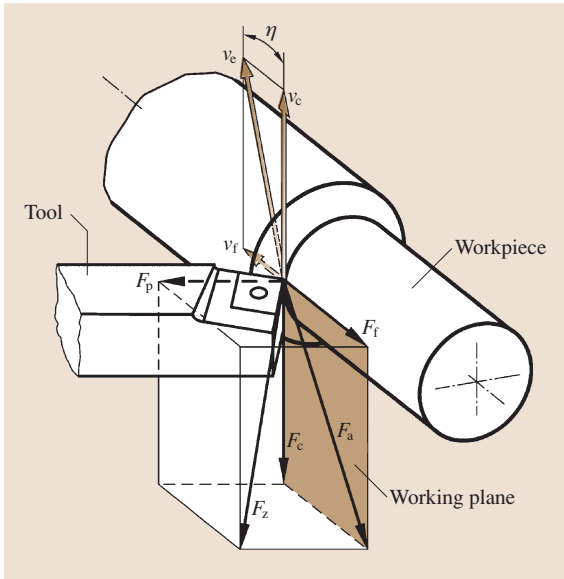


Fig. 7.138 Components of cutting force (after [7.62])

ate machine setting data such as feed rate and depth of cut [7.61].

Each material provides resistance towards the penetration of the tool during chip removal. This has to be overcome by means of a force, the machining force F. This force is analyzed by resolving it into its three components (Fig. 7.138) [7.62]. The active force F_a is formed by the cutting force F_c in the direction of the cutting motion together with the feed force F_f. The passive force F_p does not contribute to power conversion because no motion between the tool and the workpiece takes place in its direction. Thus

$$F_z = F_a + F_p = F_c + F_f + F_p \,. \tag{7.137}$$

The cutting force related to the area of the undeformed chip is defined as the specific cutting force k_c and depends on a variety of factors

$$k_c = \frac{F_c}{hb} = \frac{F_c}{a_p f} \,. \tag{7.138}$$

It is known from experiments that the specific cutting force k_c is a function of the undeformed chip thickness h. From the logarithmic representation (Fig. 7.139) [7.56] it can be seen that

$$k_c = k_{c1.1} \left(\frac{h}{h_0} \right)^{-m_c} \,. \tag{7.139}$$

Here $k_{c1.1}$ is the unit specific cutting force, i. e. k_c at $h = 1$ mm (indices 1.1 due to $k_{c1.1} = F_c/1.1$ at $b = h = 1$ mm). The quantity m_c indicates the increase and is the exponent of the specific cutting force. This equation is named in honor of the initial researcher Kienzle's cutting force formula [7.63] and can also be written as

$$F_c = k_{c1.1} b h^{1-m_c} \,. \tag{7.140}$$

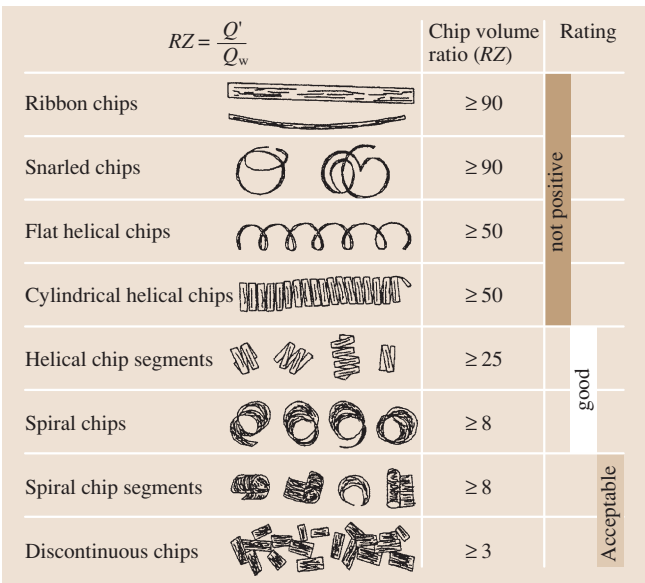


Fig. 7.137 Chip shapes, their related chip volume ratio and rating

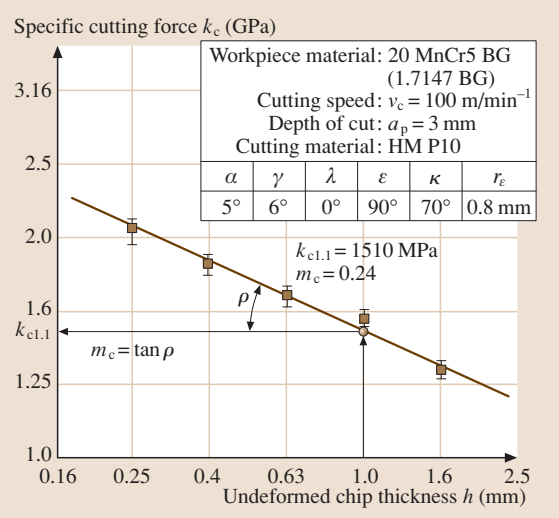


Fig. 7.139 Specific cutting force as a function of undeformed chip thickness

$k_{c1.1}$ and $1 - m_c$ are listed for various ferrous materials in Table 7.36. A direct comparison of the $k_{c1.1}$ values of different materials to indicate the machinability or the energy required for machining is not acceptable, as the exponent m_c can vary significantly. From $m_c < 1$ it follows that at a given cross-sectional area the cutting force and power requirements increase with decreasing undeformed chip thickness. This can be explained physically by the higher share of friction at decreasing undeformed chip thicknesses (see the section *Fundamentals*).

In addition to the working material and the undeformed chip thickness, k_c depends on further factors as well. Therefore additional influencing factors will be specified. The correction factors considering the cutting speed K_v, the rake angle K_γ, the cutting material K_{CM}, the tool wear K_{TW}, the cooling lubricant K_{CL}, and the workpiece shape K_{WS} are given in Table 7.37.

The passive force F_p (Fig. 7.138) [7.62], which is not relevant for the power balance due to its perpendicular orientation, is of importance for the size and form accuracy of the system – machine/workpiece/tool/fixture. Passive force F_p and feed force F_f can be merged to a thrust force F_D. For thin undeformed chip cross-sections (bh), the resultant cutting force is perpendicular to the primary cutting edge. From this it follows that

$$\frac{F_f}{F_p} = \tan\kappa \,. \tag{7.141}$$

For normal values of h and b it can be approximated that

$$F_D \approx (0.65\text{–}0.75)F_c \,, \tag{7.142}$$

wherein F_f and F_p have to be determined.

The exponential functions corresponding to the cutting force formula can be used for a more exact determination. Exponents and unit specific cutting force are given in Table 7.36.

The surface finish is determined by the profile of the cutting edge that is creating the workpiece surface and by the feed. From the shaping of the cutting edge corner radius r_ε, the theoretical maximum roughness $R_{z,th}$ according to EN ISO 4287 can be geometrically determined as

$$R_{z,th} = \frac{f^2}{8r_\varepsilon} \,. \tag{7.143}$$

This value should be regarded as a minimum for the surface roughness, which increases due to vibrations, especially at higher rotational and cutting speeds, on the formation of built-up edges (see the section *Fundamentals*), and with the progressive wear of the cutting edge.

The tool is subject to mechanical stress due to the machining force, thermal stress due to heating, and chemical attack due to the interaction of the cutting material, the workpiece material, and the surrounding medium. This results in wear on the cutting tool (see the section *Fundamentals*). Typical forms of wear are illustrated in Fig. 7.140 [7.64]. In addition, cutting edge wear, rounding of the cutting edge, and scoring may occur on the secondary cutting edge. The type of wear that determines the end of tool life (tool life criterion) is determined by the respective use. Weakening of the cutting edge by crater wear or an increase in the share of friction in the resultant force due to flank wear are critical in roughing. Cutting edge wear leads to changes in workpiece dimensions, and flank wear or scoring impair surface quality and determine the end of tool life in finishing. Often the lifetime end is set at $V_B = 0.4$ mm or $KT = 0.1$ mm. The flank is divided into three zones for more precise identification of the wear.

For a specific cutting-material–workpiece-material combination and a given tool life criterion, the tool life depends mainly on the cutting speed according to an exponential function (Taylor's equation shown on a straight line on a log–log graph) [7.65]

$$\frac{T}{T_0} = \frac{v_c^k}{C} \,. \tag{7.144}$$

Table 7.36 Values for ferrous materials $k_{c1.1}$ and $1 - m_c$

Cutting conditions								
Cutting speed		$v_c = 100\,\text{m/min}$						
Depth of cut		$a_p = 3.0\,\text{mm}$						
Cutting material		Cemented carbide P10						
Cutting edge geometry								
			α	γ	λ	ε	κ	r_ε
		Steel	5°	6°	0°	90°	70°	0.8 mm
		Cast iron	5°	2°	0°	90°	70°	0.8 mm
Material	**Material number**	R_m (N/mm²)	**Specific machining forces $k_{i1.1}$**					
			$k_{c1.1}$	$1 - m_c$	$k_{f1.1}$	$1 - m_f$	$k_{p1.1}$	$1 - m_p$
St 50-2	1.0050	559	1499	0.71	351	0.30	274	0.51
St 70-2	1.0070	824	1595	0.68	228	−0.07	152	0.10
Ck45N	1.1191 N	657	1659	0.79	521	0.51	309	0.60
Ck45V	1.1191 V	765	1584	0.74	364	0.27	282	0.57
40Mn4V	1.1157 V	755	1691	0.78	350	0.31	244	0.55
37MnSi5V	1.5122 V	892	1656	0.79	239	0.31	249	0.67
18CrNi8BG	1.5920 BG	618	1511	0.80	318	0.27	242	0.46
34CrNiMo6V	1.6582 V	1010	1686	0.82	291	0.37	284	0.72
41Cr4V	1.7035 V	961	1596	0.77	291	0.27	215	0.52
16MnCr5N	1.7131 N	500	1411	0.70	406	0.37	312	0.50
20MnCr5N	1.7147 N	588	1464	0.74	356	0.24	300	0.58
42CrMo4V	1.7225 V	1138	1773	0.83	354	0.43	252	0.49
55NiCrMoV6V	1.2713 V	1141	1595	0.71	269	0.21	198	0.34
100Cr6	1.2067	624	1726	0.72	318	0.14	362	0.47
GG30	JL1050	HB = 206	899	0.59	170	0.09	164	0.30

Table 7.37 Correction factors for cutting force calculation

Cutting speed correction factor	$K_v = \frac{2.023}{v_c^{0.153}}$ for $v_c < 100\,\text{m/min}$ $K_v = \frac{1.380}{v_c^{0.07}}$ for $v_c > 100\,\text{m/min}$
Rake angle correction factor	$K_\gamma = 1.09–0.015\angle°$ (steel) $K_\gamma = 1.03–0.015\angle°$ (cast iron)
Cutting material correction factor	$K_{CM} = 1.05$ (HSS) $K_{CM} = 1.0$ (cemented carbide) $K_{CM} = 0.9–0.95$ (ceramic)
Tool wear correction factor	$K_{TW} = 1.3–1.5$ $K_{TW} = 1.0$ for sharp cutting edge
Cutting fluid correction factor	$K_{CL} = 1.0$ (dry) $K_{CL} = 0.85$ (non-water soluble coolant) $K_{CL} = 0.9$ (emulsion-type coolant)
Workpiece shape correction factor	$K_{WS} = 1.0$ (outer diameter turning) $K_{WS} = 1.2$ (inner diameter turning)

Here, T_0 and v_c are reference values. T_0 is normally set at $T_0 = 1$ min. C is the cutting speed for an operating period of $T_0 = 1$ min.

The Taylor straight line is plotted on the basis of a wear/tool life turning test according to ISO 3685. With these tests suitable settings for high-speed steel,

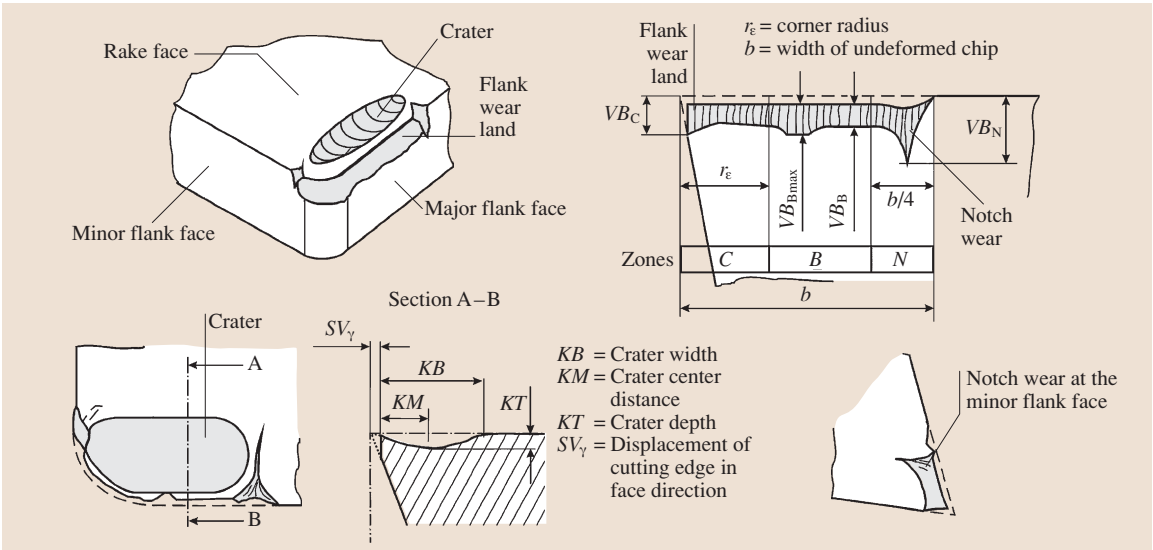


Fig. 7.140 Forms of wear in turning (ISO 3685)

hard metals of all machining application categories, and ceramic inserts have been determined. It is usually sufficient to determine the width of the flank wear land VB and/or the crater depth KT, as well as the distance from the crater center to the point of the cutting tip. Table 7.38 normal values of the gradient exponent k and the cutting speed C for a tool life of $T = 1$ min and a width of wear land of $VB = 0.4$ mm shows for various materials.

For metal cutting machines the optimum cutting speed must be established to meet economical criteria (Fig. 7.141) [7.56]. The optimum cutting speed in relation to time is expressed by

$$v_{c,opt} = C(-k-1)t_{CT}^{1/k} \,. \qquad (7.145)$$

Optimization of the cutting speed to minimize unit costs takes into account not only the tool changing time t_{CT} but also the tool costs per cutting edge K_{CT} and the

Table 7.38 Coefficients of Taylor tool life equations

Taylor-function $v_c = CT^{1/k}$		Tungsten carbide uncoated		coated		Oxide ceramic (St) nitride ceramic (GG)	
Material	Material number	C (m/min)	k	C (m/min)	k	C (m/min)	k
St 50-2	1.0050	299	−3.85	385	−4.55	1210	−2.27
ST 70-2	1.0070	226	−4.55	306	−5.26	1040	−2.27
Ck45N	1.1191N	299	−3.85	385	−4.55	1210	−2.27
16MnCrS5BG	1.7131BG	478	−3.13	588	−3.57	1780	−2.13
20MnCr5BG	1.7147BG	478	−3.13	588	−3.57	1780	−2.13
42CrMoS4V	1.7225V	177	−5.26	234	−6.25	830	−2.44
X155CrVMo12-1	1.2379	110	−7.69	163	−8.33	570	−2.63
X40CrMoV5-1G	1.2344	177	−5.26	234	−6.25	830	−2.44
GG-30	GJL-300	97	−6.25	184	−6.25	2120	−2.50
GG-40	GJL-400	53	−10.00	102	−10.0	1275	−2.78

Table values are valid for $a_p = 1$ mm, $f = 1$ mm, $VB = 0.4$ mm

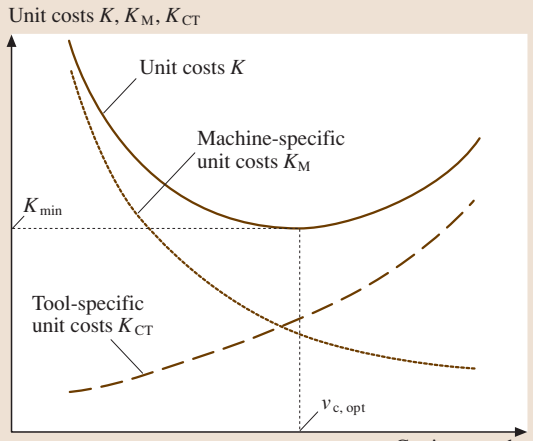


Fig. 7.141 Manufacturing costs as a function of cutting speed v_c

hourly machine tool rate K_M

$$v_{c,opt} = C(-k-1)\left(t_{CT} + \frac{K_{CT}}{K_M}\right)^{\frac{1}{k}} . \qquad (7.146)$$

Drilling and Reaming

Drilling is a metal cutting process with a rotary cutting motion (primary motion). The tool, i. e. the drill, performs the feed motion in the direction of the rotational axis. Figure 7.142 [7.55] shows common drilling processes. With drilling into solid metal either through holes or blind holes can be produced. The tool is usually a twist drill. Drilled holes are either enlarged with twist drills or with countersinks or counterbores having two or more cutting edges. Step drills produce stepped holes. They usually have multiple cutting edges; for manufacturing reasons, not every cutting edge has to support all parts of the contour (e.g. one cutting edge may break the edge of a step, while the adjacent one produces a planar surface). Centre drills are special profile drills with a thinner spigot and a short, stiff drill section in order to attain a good centering effect. Trepanning cutters produce an annular cut into solid material with the coincidental formation of a solid cylinder or plug. Taps are used to cut internal threads. Reaming is a hole enlarging process with a small undeformed chip thickness, for producing holes of precise size and shape with a high-quality surface.

For drilling holes with diameters of 1–20 mm and with drilling depths up to five times the diameter, the twist drill is the tool most commonly used (Fig. 7.143) [7.55]. The twist drill consists of the shank and the cutting section. The shank is used to locate and clamp the drill in the machine tool. It is straight or tapered. If a high cutting torque is to be transmitted, tangential flat surfaces transmit the force. The cutting section has a complex geometry, which can be modified to adapt the drill to the respective machining task. Essential parameters are the profile and the web thickness, the flute geometry and the helix angle, i. e. the pitch of the flutes, the ground drill-point, and the drill-point angle. Especially the ground drill-point and the drill-point angle can be determined by the user. The profile of the twist drill is designed in such a way that the flutes provide the maximum space for chip removals while ensuring that the drill is capable to adequately withstand torsional stresses. These are the two main requirements. A further de-

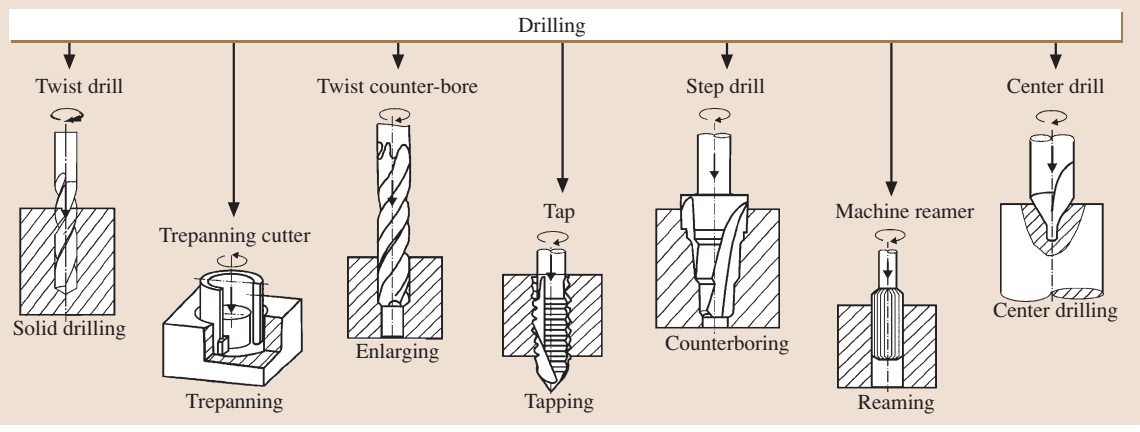


Fig. 7.142 Drilling processes (DIN 8589)

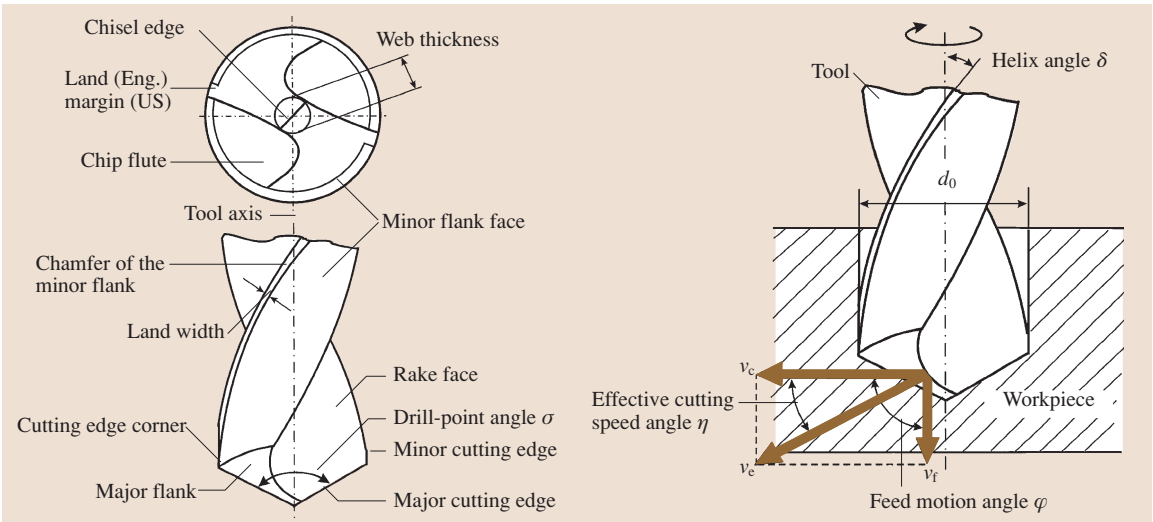


Fig. 7.143 Terminology and mode of operation of twist drills (DIN 8589)

mand on the drill may be the production of favorable chip forms. This has led to a diversity of special profiles and allows the adaption of the drilling process to particular constraints. Working material also has to be removed from the area in front of the core of the twist drill; with the special boundary condition that the cutting speed v_c in the center is zero. This is achieved by the chisel edge, which connects the major cutting edges.

Along the major and minor cutting edges, the rake angle γ as an important factor influencing the drilling process is not constant but already decreases in front of the major cutting edge from the outside towards the inside. In Fig. 7.144 [7.56], the rake angles at three positions on the major cutting edge and chisel edge, respectively, are shown.

At the outer diameter the rake angle is almost identical to the helix angle δ and decreases in direct proportion to the diameter. Close to the chisel edge, the rake angles on the major cutting edge are already strongly negative. Here the workpiece material has to be displaced in radial direction. As shown in the left

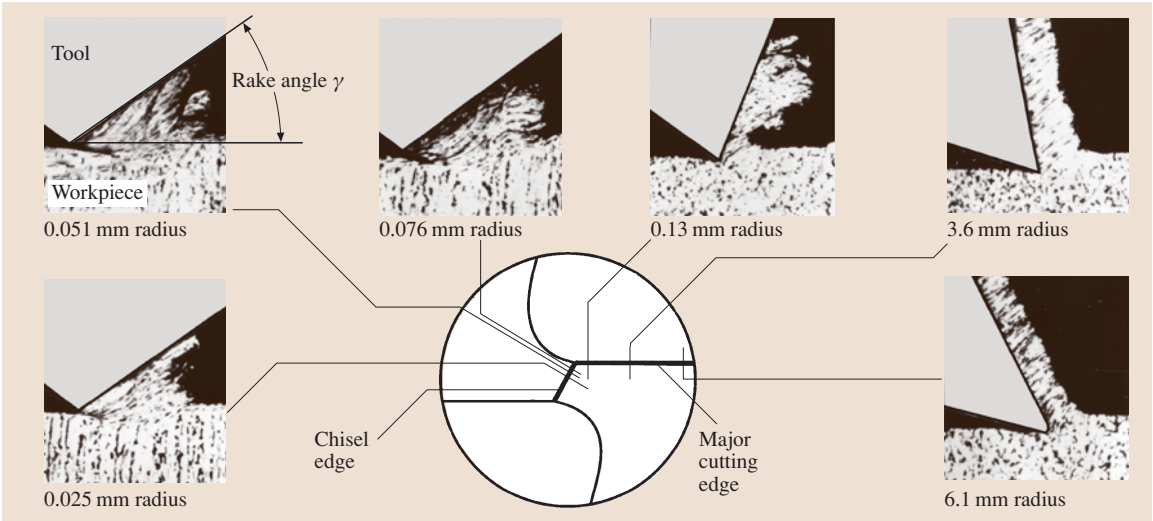


Fig. 7.144 Rake angles at different position on a drill tip

of Fig. 7.144 all rake angles on the chisel edge are negative. These negative rake angles and the material displacement effect generate high pressures in the area of the chisel edge. To diminish this effect, twist drills are pointed. The center of the drill is tapered by profile grinding in the direction of the flute and towards the drill tip on a conical or similar surface. In this way, the rake angle at the chisel edge is increased and/or the chisel edge is shortened.

The most important type of wear concerning the twist drill is flank wear at the corners. This wear, which is mainly caused by abrasion, generates an increase in the torsional load on the drill, as higher machining forces are present in the corner area. This torsional load may cause the drill to break. Worn twist drills are therefore reground until the damaged area of the secondary cutting edge is removed.

Cutting Forces. The forces and moments in drilling are calculated on the basis of Kienzle's approach [7.63, 66]. Figure 7.145 [7.56] illustrates the cutting geometry and the forces implied in drilling processes. The forces arising per cutting edge, which are assumed to act in the middle of the cutting edge, are divided into their components F_c, F_p and F_f. The cutting forces F_{c1} and F_{c2} generate the cutting moment via the lever arm r_c, the cutting moment M_c

$$M_c = (F_{c1} + F_{c2}) r_c \,, \tag{7.147}$$

$$F_{c1} = F_{c2} = F_{cZ} \,, \tag{7.148}$$

$$M_c = F_{cZ} 2 r_c \,. \tag{7.149}$$

F_f is the sum of the feed forces F_{f1} and F_{f2},

$$F_f = F_{f1} + F_{f2} \,, \tag{7.150}$$

$$F_{f1} = F_{f2} = F_{fZ} \,, \tag{7.151}$$

$$F_f = 2 F_{fZ} \,. \tag{7.152}$$

In the ideal case, i. e. with a symmetrical drill, the passive forces F_{p1} and F_{p2} cancel each other out. If there are symmetry deviations, F_{p1} and F_{p2} generate interference forces, which impair the quality of the hole. The cutting force per cutting edge results in

$$F_{cZ} = b h^{(1-m_c)} k_{c1.1} \,, \tag{7.153}$$

$$h = F_z \sin\kappa \,, \tag{7.154}$$

$$b = \frac{d_0 - d_i}{2 \sin\kappa} \,. \tag{7.155}$$

By analogy, the feed force is

$$F_f = b h^{(1-m_c)} k_{f1.1} \,. \tag{7.156}$$

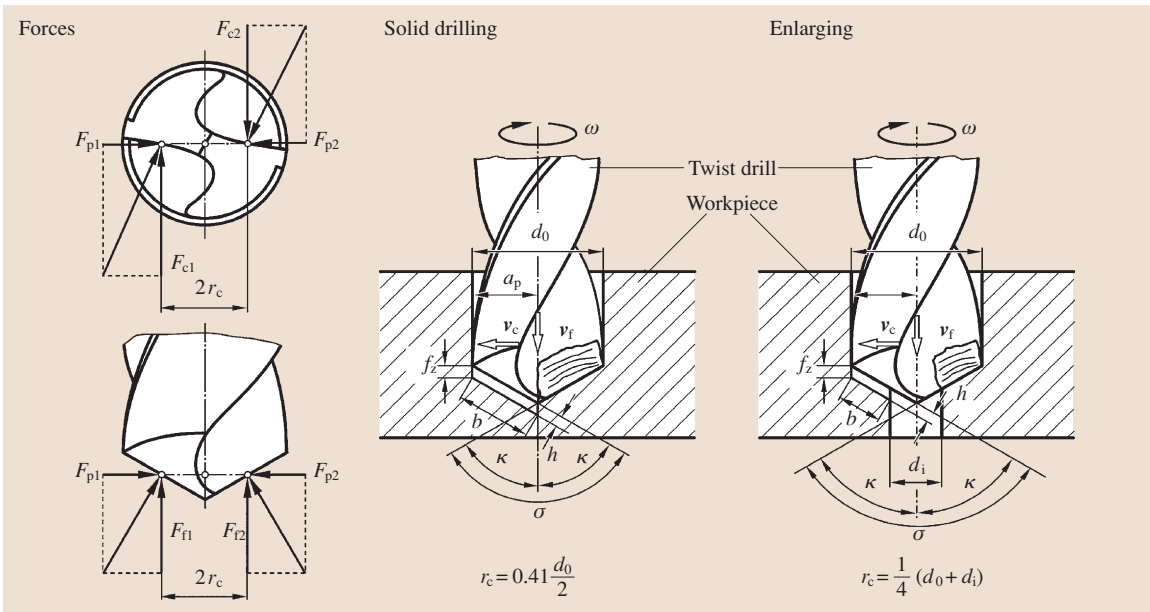


Fig. 7.145 Cutting geometry and machining forces in drilling

Table 7.39 Cutting force components for drilling

Material	Mat. No.	R_m (N/mm^2)	$1-m_c$	$k_{c1.1}$ (N/mm^2)	$1-m_f$	$k_{f1.1}$ (N/mm^2)
18CrNi8	1.5920	600	0.82 ± 0.04	2690 ± 230	0.55 ± 0.06	1240 ± 160
42CrMo4	1.7225	1080	0.86 ± 0.06	2720 ± 420	0.71 ± 0.04	2370 ± 230
100Cr6	1.2076	710	0.76 ± 0.03	2780 ± 220	0.56 ± 0.07	1630 ± 300
46MnSi4	1.5121	650	0.85 ± 0.04	2390 ± 250	0.62 ± 0.02	1360 ± 100
Ck60	1.1221	850	0.87 ± 0.03	2200 ± 200	0.57 ± 0.03	1170 ± 100
St50	1.0531	560	0.82 ± 0.03	1960 ± 160	0.71 ± 0.02	1250 ± 70
16MnCr5	1.7131	560	0.83 ± 0.03	2020 ± 200	0.64 ± 0.03	1220 ± 120
34CrMo4	1.7220	610	0.80 ± 0.03	1840 ± 150	0.64 ± 0.03	1460 ± 140
Grey cast iron						
Up to G-22	–	–	0.51	504	0.56	356
Over G-22	–	–	0.48	535	0.53	381

Values are given in Table 7.39. The feed forces are strongly dependent on the shape of the chisel edge. They can be lowered significantly by web thinning [7.56]. Wear causes them to reach twice their original value or more.

Surface quality in drilling with twist drills corresponds to roughing with $R_z = 10-20\,\mu m$. The surface roughness can be reduced by reaming with increased dimensional accuracy. The application of solid cemented carbide drills provides another solution. When drilling solid metal, surface qualities, dimensional accuracy, and accuracy of shape like those obtained with reaming are achieved. Most of the drilling tools are further improved by suitable coatings.

Short-Hole Drilling. Short-hole drilling with drilling depths of $L < 2D$ covers a large proportion of bolt hole drilling, through hole drilling and tapping. For this, short-hole drills with indexable inserts may be used for diameters from 10 to over 120 mm. Their advantage compared with twist drills is the absence of a chisel edge, and the increase in cutting speed and feed rate achieved with indexable cemented carbide or ceramic inserts. Due to the asymmetrical machining forces, the use of short-hole drills requires rigid tool spindles similar to those found on common machining centers and milling machines. The higher rigidity of the tool enables pilot drilling of inclined or curved surfaces with accuracy of IT7.

Milling

Classification of Milling Processes. In milling, the necessary relative motion between the tool and the workpiece is achieved by means of a circular cutting motion of the tool and a feed motion perpendicular to or at an angle to the axis of rotation of the tool. The cutting edge is not continuously in engagement with the workpiece. Therefore, it is subject to alternating thermal and mechanical stresses. The complete machine-tool–workpiece-fixture system is dynamically stressed by the interrupted cutting action.

Milling processes are classified according to DIN 8589 on the basis of the following:

- The nature of the resulting workpiece surface
- The kinematics of the cutting operation
- The profile of the milling cutter

Milling can be used to produce a practically infinite variety of workpiece surfaces. A distinguishing feature of a process is the cutting edge (major or minor) that produces the workpiece surface (Fig. 7.146): in face milling the minor cutting edge is located at the face of the milling cutter, while in peripheral milling the major cutting edge is located on the circumference of the milling cutter.

A distinction can be made on the basis of the feed direction angle φ (Fig. 7.147): in down-milling the feed direction angle φ is $> 90°$, thus the cutting edge of the milling cutter enters the workpiece at the maximum undeformed chip thickness, while in up-milling the feed direction angle φ is $< 90°$, thus the cutting edge enters at the theoretical undeformed chip thickness $h = 0$. This initially results in pinching and rubbing.

A milling operation may include both up-milling and down-milling. The principal milling processes are summarized in Fig. 7.148.

Plain Face Milling with End Milling Cutters. The kinematics of cutting and the relationship of the cutting

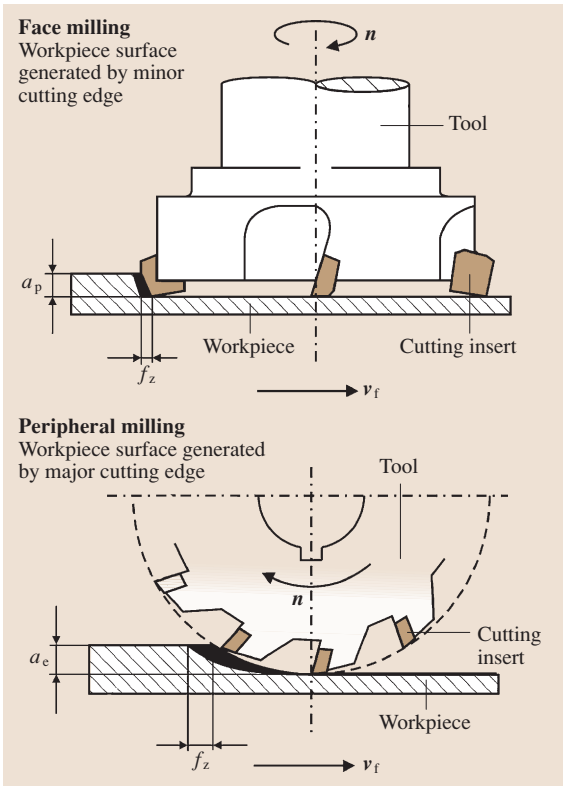


Fig. 7.146 Comparison of face milling and peripheral milling

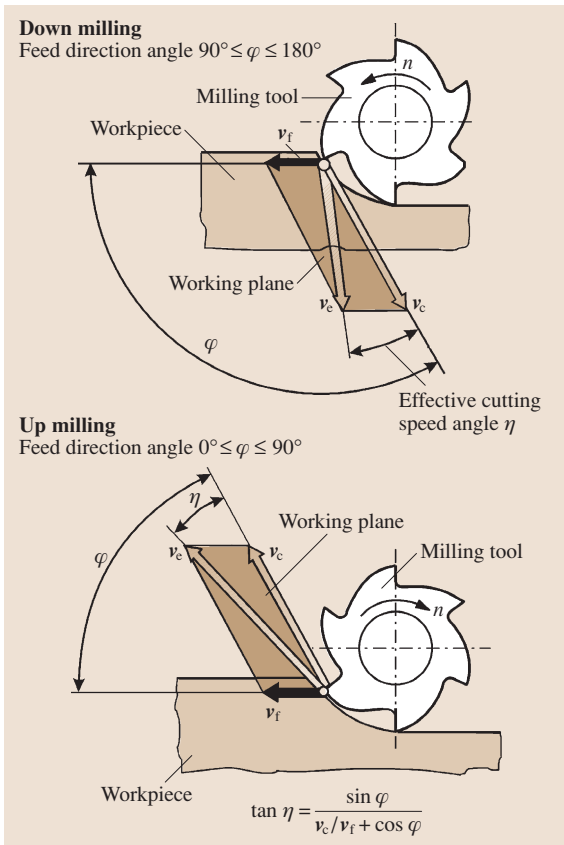


Fig. 7.147 Comparison of down and up milling (DIN 6580E)

forces during milling will be discussed with reference to plain face milling with an end milling cutter.

Kinematics of Cutting in Milling. To describe the process, it is necessary to distinguish between the tool-workpiece engagement variables and the cutting variables. The engagement variables, which are expressed in relation to the working plane, describe the interaction of the cutting edge and the workpiece. The working plane is described by the cutting speed vector v_c and the feed velocity vector v_f. In milling the engagement variables are (Fig. 7.149): depth of cut a_p, measured at a right angle to the working plane, cutting engagement a_e, measured in the working plane at right angles to the feed direction, and feed of the cutting edge f_z, measured in the feed direction.

For a full description of the kinematics of cutting, the following data are required: milling cutter diameter D, number of teeth z, tool excess x_e, and the cutting edge geometry (side rake angle γ_f, back rake angle γ_p, side clearance angle α_f, back clearance angle α_p, entry angle κ_r, angle of inclination λ_s, cutting edge radius r, and chamfer).

Due to the interrupted cut, the entry and exit conditions of the cutting edge, and the types of contact, are especially important for milling processes. The types of contact describe the nature of the first and last contacts of the cutting edge with the workpiece. They can be determined from the entry angle φ_E, the exit angle φ_A, and the tool geometry. Generating the first point of contact with the cutting edge tip should be avoided.

From the engagement variables the cutting variables can be derived, which indicate the dimensions of the layer of material to be removed from the workpiece. The cutting variables are not identical to the chip variables, which describe the dimensions of the resulting chips. The cutting edges describe cycloids in relation to the workpiece. As the cutting speed is significantly higher than the feed velocity, they can be approximated by cir-

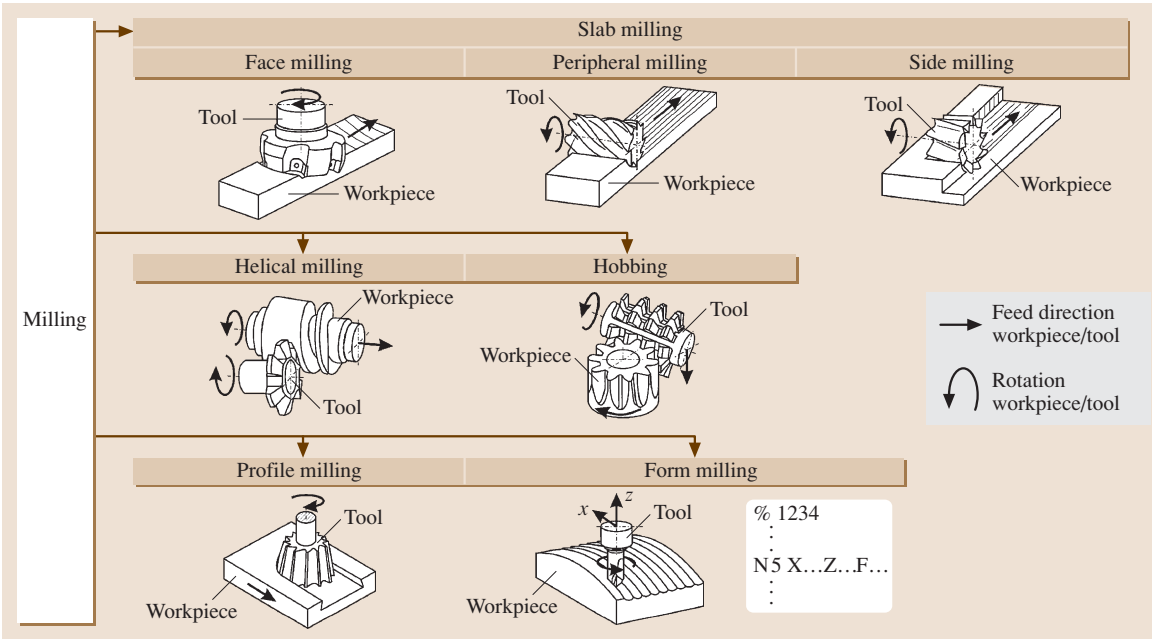


Fig. 7.148 Milling processes (DIN 8589)

cular paths. Taking this into account, the undeformed chip thickness is (Fig. 7.149)

$$h(\varphi) = f_z \sin\kappa \sin\varphi \,. \tag{7.157}$$

With the undeformed chip width $b = a_p / \sin\kappa$, the undeformed chip cross-section is

$$A(\varphi) = bh(\varphi) = a_p f_z \sin\varphi \,. \tag{7.158}$$

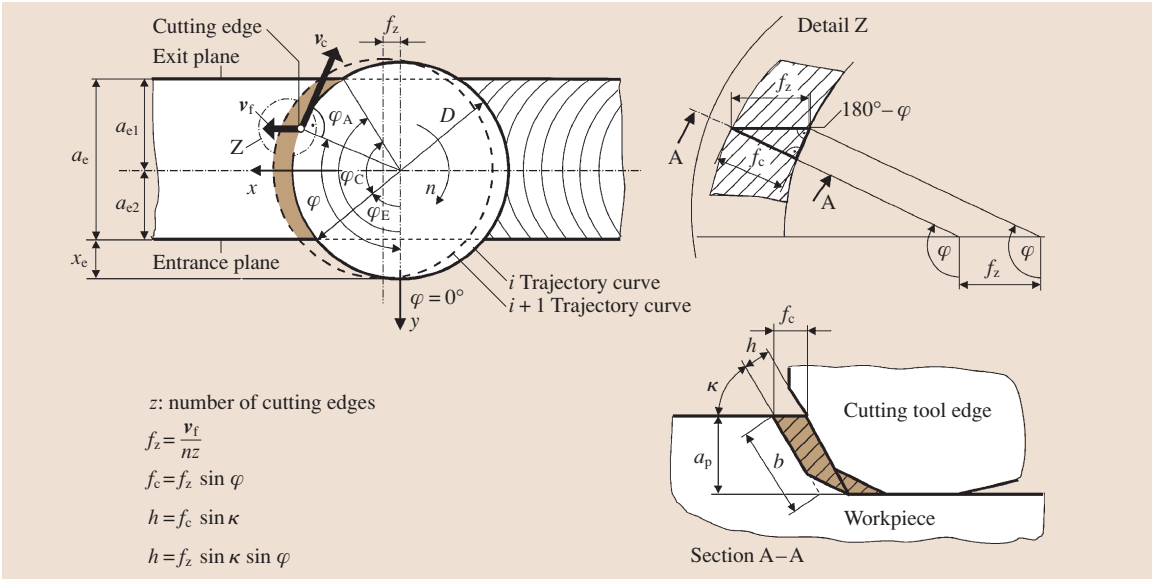


Fig. 7.149 Tool–workpiece engagement variables in plain face milling

The time-related chip volume is $Q = a_c a_p v_f$. The undeformed chip thickness is a function of the entry angle φ and is thus not constant as in turning. The evaluation of the milling process is based on the mean undeformed chip thickness

$$h_m = \frac{1}{\varphi_c}\int_{\varphi_E}^{\varphi_A} h(\varphi)\,\mathrm{d}\varphi = \frac{1}{\varphi_c} f_z \sin\kappa(\cos\varphi_E - \cos\varphi_A)\,. \qquad (7.159)$$

Cutting Force Components. The machining force required for chip formation has to be absorbed by the cutting edge and the workpiece. According to DIN 6584, the machining force F can be resolved into an active force F_a, which lies in the working plane, and a passive force F_p, which is perpendicular to the working plane. The direction of the active force F_a changes with the entry angle φ. The components of the active force can be expressed in relation to the following directions (Fig. 7.150).

- Direction of cutting speed v_c: the components cutting force F_c and perpendicular cutting force F_{cN} relate to a co-rotating system of coordinates (tool-specific components of the active force).
- Direction of rate of feed v_f: the components feed force F_f and perpendicular feed force F_{fN} relate to a fixed system of coordinates (workpiece-specific components of the active force).

For converting the active force from the fixed system of coordinates into a co-rotating system, the following equations apply

$$F_c(\varphi) = F_f(\varphi)\cos\varphi + F_{fN}(\varphi)\sin\varphi\,, \qquad (7.160)$$
$$F_{cN}(\varphi) = F_f(\varphi)\cos\varphi + F_{fN}(\varphi)\cos\varphi\,, \qquad (7.161)$$
$$F_x(\varphi) = F_f(\varphi)\,, \qquad (7.162)$$
$$F_y(\varphi) = F_{fN}(\varphi)\,. \qquad (7.163)$$

This transformation is important if, for instance, the cutting force F_c is to be measured with a three-component force-measuring dynamometer that holds the workpiece. Figure 7.150 shows the pattern of the components of the active force in the tool- and workpiece-specific systems of coordinates for plain face milling with an end milling cutter.

Prediction of Cutting Force Components. Kienzle's machining force equation [7.63] can also be applied to milling. For the machining force components cutting force F_c, perpendicular cutting force F_{cN} and passive force F_p the calculation is

$$F_i = Ak_i\,, \qquad (7.164)$$

where $i = \mathrm{c}, \mathrm{cN}, \mathrm{p}$. A is the undeformed chip cross-section and k_i is the specific machining force. Owing to the wide range of undeformed chip thicknesses that is covered by milling (the undeformed chip thickness depends on the engagement angle φ), Kienzle's relationship only applies to certain areas. The undeformed chip thickness range of $0.001\,\mathrm{mm} < h < 1.0\,\mathrm{mm}$ is divided into three sections (Fig. 7.151). For each section,

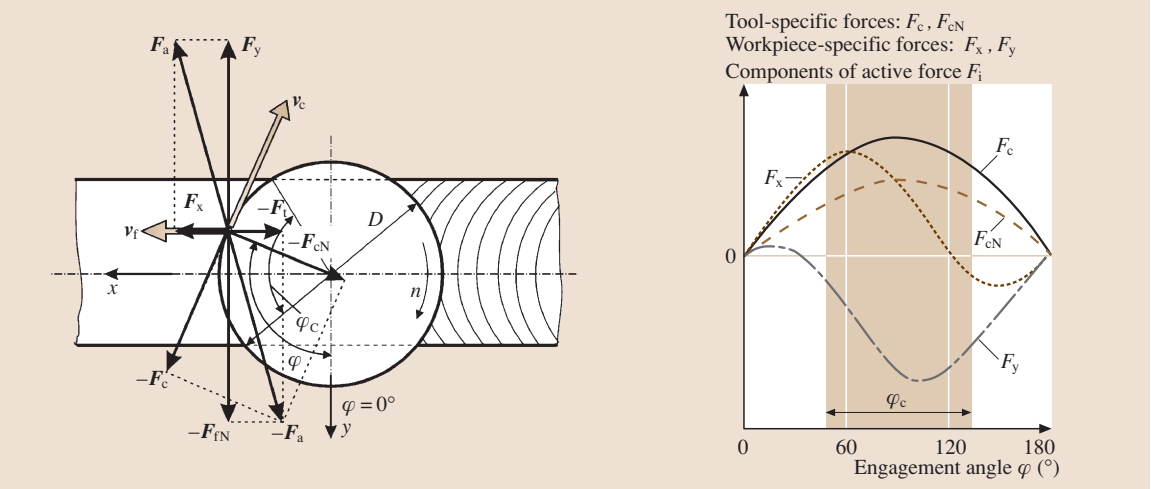


Fig. 7.150 Components of cutting force in plain face milling

a straight line can be determined, depending on, the unit specific cutting force k_i and its exponent m_i. For the specific cutting force the following equations apply

$$k_i = k_{i1.0,01} h^{-m_{i0,01}} \quad \text{for} \quad 0.001\,\text{mm} < h < 0.01\,\text{mm}\,, \tag{7.165}$$

$$k_i = k_{i1.0,1} h^{-m_{i0,1}} \quad \text{for} \quad 0.01\,\text{mm} < h < 0.1\,\text{mm}\,, \tag{7.166}$$

$$k_i = k_{i1.1} h^{-m_i} \quad \text{for} \quad 0.1\,\text{mm} < h < 1\,\text{mm}\,, \tag{7.167}$$

where $i = \text{c, cN, p}$.

Thus the cutting force for milling with an end milling cutter is

$$F_i = b k_{i1.1} h^{1-m_i}\,, \tag{7.168}$$

where $i = \text{c, cN, p}$.

The corresponding component of the cutting force can be calculated for milling if unit specific cutting force k_i and its exponent m_i for the workpiece/cutting material combination and the cutting conditions are available. The machining indices for axial plane face milling with a milling head are given in Table 7.40 for a number of workpiece materials and cutting conditions [7.63]. Often, though, to estimate the cutting force during milling, values obtained from turning will have to be used.

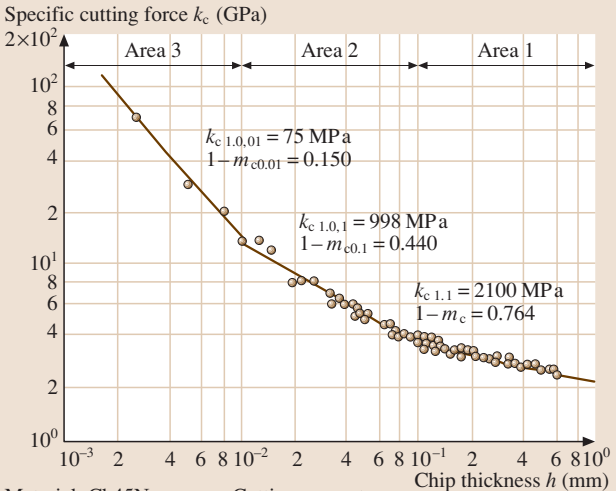
The milling machine capacity is designed on the basis of the average machining force

$$F_i = b k_{i1.1} h_m^{1-m_i} K_{\text{pro}} K_\gamma K_\text{v} K_{\text{TW}} K_{\text{CM}}\,, \tag{7.169}$$

where $i = \text{c, cN, p}$.

In this equation, h_m is the mean undeformed chip thickness, $K_{\text{pro}} = 1.2\text{–}1.4$ is the correction factor for the manufacturing process (the factor takes into account the fact that the machining indices were obtained from turning tests), K_γ is the correction factor for the rake angle (see the section *Turning*), K_v is the correction factor for the cutting speed, K_{TW} is the correction factor for tool wear, and K_{CM} is the correction factor for the cutting material. Experimental research on plain face milling has proven that the influence of wear on the machining force components cannot be ignored.

Vibrations. Depending upon the elasticity frequency response of the complete milling machine-milling cutter-workpiece-fixture system, the metal cutting forces generate vibrations that may affect surface quality and tool life. According to their origin, these vibrations are divided into separately excited and self-excited vibrations [7.67, 68].



Material: Ck45N
Material-No. 1.1191N
Cutting tool: HM P 25
Cutting conditions:
$v_c = 190$ m/min

Cutting geometry

γ_r	γ_a	α_r	α_a	λ_s	κ
–4°	–7°	6°	23°	–6°	75°

ε	Chamfer
90°	45° × 0.06 mm

Fig. 7.151 Specific cutting force in plain face milling

In the case of separate excitation, the complete system vibrates at the frequency of the exciting external forces. Due to the intermittent cutting action of milling the cutting edges are not constantly in engagement. With a multiple-edged milling cutter, the number of cutting edges in engagement at any time should be taken into account. Depending on the ratio a_c/D, z_{iE} cutting edges are in engagement, the following relationship being valid

$$z_{\text{iE}} = \frac{\varphi_\text{c}}{2\pi} z\,, \quad \text{where} \quad \frac{\varphi_\text{c}}{2} = \frac{a_\text{c}}{D}\,. \tag{7.170}$$

The mean cutting force acting on the milling cutter and thus on the spindle of the milling machine is

$$F_{\text{cm}} = z_{\text{iE}} F_{\text{cmz}}\,, \tag{7.171}$$

where F_{cmz} is the mean cutting force of a single cutting edge. Superimposed on the mean cutting force is a dynamic force element. The larger the value of z_{iE} the smaller the force amplitude; if z_{iE} is an integer, the cutting force amplitude is at a minimum. The dynamic force element leads to separately excited vibrations between the workpiece and the milling cutter.

Table 7.40 Cutting force components for plain face milling

Cutting edge geometry							Main and incremental values for spec. machining force in axial face milling						
γ_f	γ_p	α_f	α_p	λ_s	κ_r	κ_F	ε	Chamfer					
Material number	Cutting material	v_c (m/min)	Cutting edge geometry	$k_{c1.1}$ (N/mm²)	m_c	$K_{cN1.1}$ (N/mm²)	m_{cN}	$k_{p1.1}$ (N/mm²)	m_p				
Neg.	−7°	6°	23°	−6°	75°	60°/30°/0°	90°	1.4/0.8/1.4					
Pos.	8°	9°	29°	8°	75°	45°/0°	90°	0.8/1.4					
St 52-3 N	1.0570 N	HM P25	120	Neg.	1831	0.29	809	0.54	705	0.41			
				Pos.	1469	0.25	447	0.57	174	0.56			
Ck45N	1.1191 N	HM P25	190	Neg.	1506	0.45	708	0.62	653	0.52			
				Pos.	1533	0.29	497	0.70	164	0.77			
X22CrMoV12-1	1.4923	HM P40	120										

In the case of self-excitation, the complete machining system vibrates at one or more eigenfrequencies, without an external interference force affecting the system. Special importance is attached to self-excited vibrations that arise because of the regenerative effect which are also referred to as *regenerative chatter*. The chatter is caused by variations in cutting force due to changes in undeformed chip thickness [7.67]. Chatter can be influenced by varying the cutting speed, depth of cut, feed rate, and cutting edge geometry.

Tool Wear in Milling. Due to the intermittent cutting action in milling, the cutting material is subject to alternating thermal and mechanical stresses. As a consequence, not only face and flank wear but also cracking in the cutting tip may determine tool life. Figure 7.152 depicts the tool life travel per tooth based on the flank wear of the primary cutting edge of an end ball nose milling cutter for different cutting tool materials [7.56]. As shown especially with the development of cubic boron nitride (CBN), the wear behavior of milling tools could be significantly improved, thus allowing finish-machining of hardened materials by milling [7.69–71]. Suitable data for cutting parameters are accessible via catalogues of all major milling tool manufacturers. Depending on the cutting conditions, surface roughness values comparable to those obtained in grinding are achieved. In grinding, the accuracy of shape is achieved by spark-out. As there has to be a minimum undeformed chip thickness in milling, shape defects occur, which

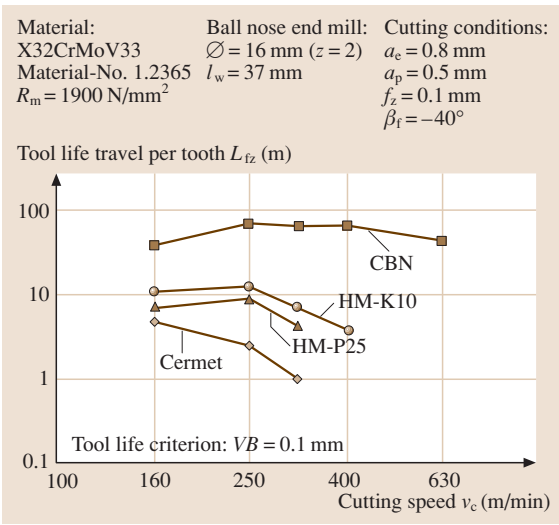


Fig. 7.152 Milling tool wear depending on cutting tool material and speed (after [7.56])

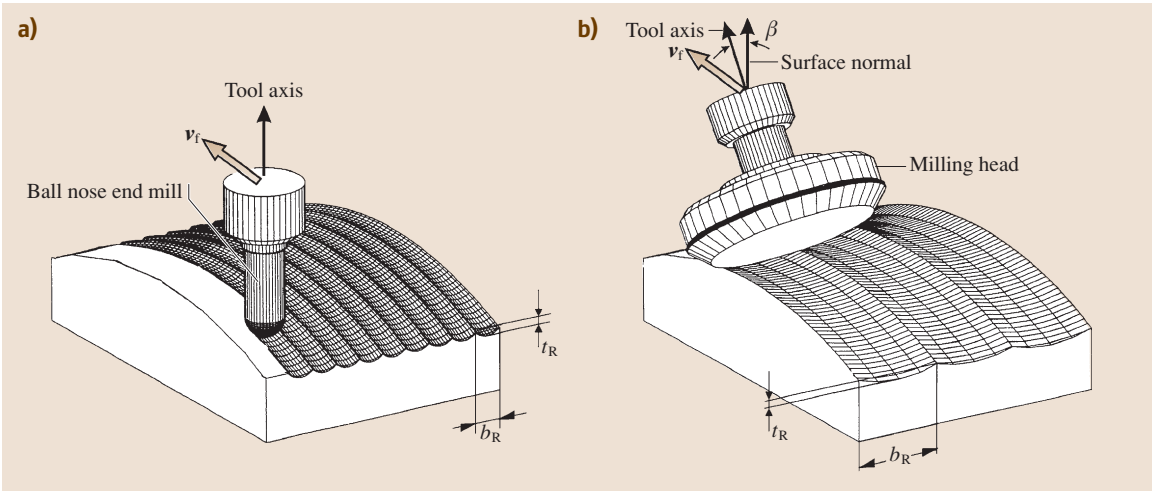


Fig. 7.153a,b Form milling by (a) 3-axis milling and (b) 5-axis milling

can be attributed to the following influencing variables: environment, operating behavior of the milling machine, inhomogeneity in material hardness, heat absorption of the workpiece due to metal cutting, and a change of residual stresses in the edge zone of the workpiece.

Form Milling. Hollow-mold tools such as deep-drawing tools are manufactured by both chip-removal and chipless machining processes. Milling plays a dominant role as a controlled shaping process. The essential characteristic in form milling is the number of actively controlled axes, a distinction accordingly being made between three-axis milling and five-axis milling (Fig. 7.153). In five-axis milling, not only the tip but also the axial direction of the milling cutter is continuously and simultaneously controlled relative to the workpiece coordinate system. The profile of the milled grooves determines the productivity and quality of the process (small final finishing with a small profile depth being desired). It is formed by machining a curved surface in parallel lines and depends on the milling cutter geometry, the workpiece geometry, and the method of work. For a given groove depth t_R, five-axis milling with an end milling cutter produces significantly larger groove widths b_R than three-axis milling with a convex ball nose end milling cutter.

Other Processes: Planing and Shaping, Broaching, Sawing

Planing and Shaping. According to a German standard a distinction can be made between planing and shaping. Chip removal is accomplished during the working stroke with a single-point cutting tool. The following return stroke resets the tool to its original position. The feed is performed intermittently, usually at the end of a return stroke.

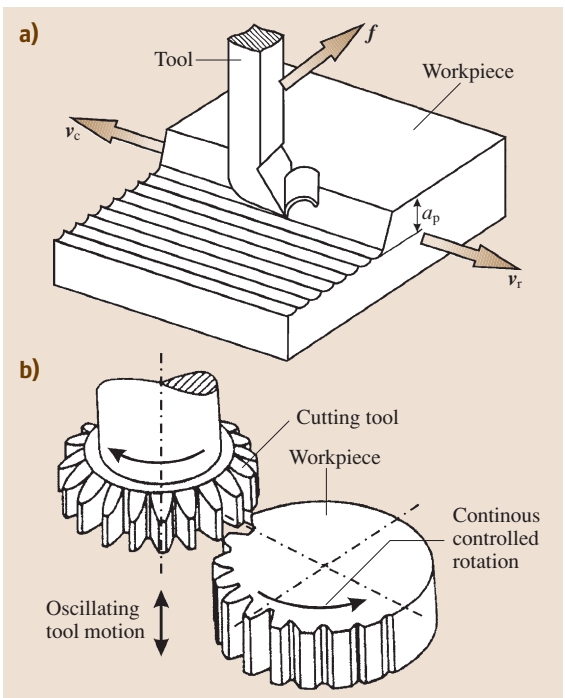


Fig. 7.154 (a) Planing and (b) gear shaping

In planing, the workpiece performs the cutting motion with the cutting speed v_c and the return motion with v_r. Feed f and depth of cut a_p are accomplished by the tool (Fig. 7.154a). In shaping, the tool performs the cutting and return motion, while feed and engagement are accomplished by the workpiece or the tool.

The reciprocating motion of the workpiece (in planing) or the tool (in shaping) produces high inertia forces and limits the cutting speed. As a guideline for the cutting speed, for machining steels the ranges $v_c = 60–80$ m/min (roughing) and $v_c = 70–100$ m/min (finishing) are well established for cemented carbide tools. Frequently used special methods are manufacturing by planing or shaping (Fig. 7.154b) to produce involute gears [7.72].

Broaching. In broaching material is removed using a multiple-pointed tool, the teeth of which are one behind the other and are successively stepped by one layer of the material to be removed. Thus no feed motion is required, as it is *built into* the tool. The cutting motion is usually linear or, under certain circumstances, helical or circular.

The advantages of the process are the high productivity and the possibility of finishing workpieces with a single tool because high surface finish and dimensional accuracies with tolerances up to IT7 can be achieved. Due to the high cost of the tools, the main areas of use are series and mass production; a new tool is required for each workpiece shape.

A basic distinction is made between internal broaching and external broaching. In internal broaching, the broach is pushed or pulled through a hole, while in external broaching the tool moves along the surface.

Broaches are divided into roughing, finishing, and calibrating teeth sections (Fig. 7.155). Normal undeformed chip thicknesses in flat broaching of steels are $h_z = 0.01–0.15$ mm for roughing and $h_z = 0.003–0.023$ mm for finishing. Broaching of cast materials is carried out up to a thickness of $h_z = 0.02–0.2$ mm in the roughing section and $h_z = 0.01–0.04$ mm in the finishing section.

Cutting speeds are restricted by the hardness of the chosen cutting material at high temperature and by the efficiency of the machine. The cutting material most commonly used, high-speed steel (HSS), permits only low cutting speeds owing to its hot hardness of approximately 600 °C. Cutting speeds of $v_c = 1$ to 30 m/min are used, with speeds of up to 70 m/min in exceptional cases. The capacity of the process can be increased by using TiN-coated HSS or cemented carbide. Due to these developments broaching of case-hardened steel workpieces with cemented carbide tools has also become possible [7.73]. High cutting speeds require high drive power outputs to accelerate and decelerate the tool and the broach slide, causing a disproportionate increase in equipment costs. Vibration problems might also occur, especially with thin internal broaches.

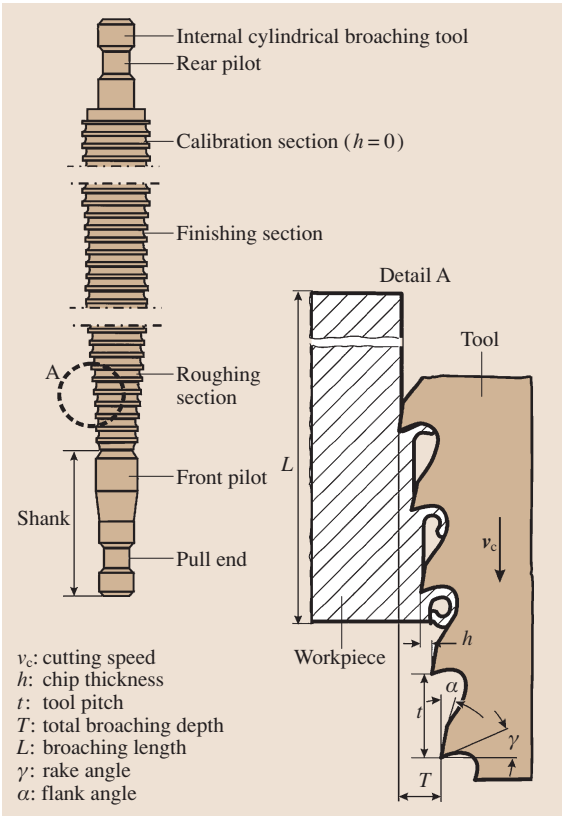


Fig. 7.155 Internal cylindrical broaching tool and process characteristics

In broaching, mainly mineral oils are used as cutting fluid for lubricating and cooling in the contact zone, but above all to prevent the formation of built-up edges and to carry away the chips. They usually contain EP (extreme pressure) additives, which are nowadays mostly chlorine-free.

Besides the above-mentioned internal and external broaching operations also special process configurations like gear scraping or turn broaching are used in industry [7.72].

Sawing. Sawing is metal cutting with a multiple-pointed tool having a small width of cut for severing

or slitting workpieces. The rotational or linear principal motion is performed by the tool. The teeth of the tool are offset in alternate directions. Thus the kerf is widened in relation to the saw blade, thus reducing friction between tool and workpiece (Fig. 7.156).

Band sawing is sawing with a continuous, usually straight cutting motion of a rotating endless band. The motions and cutting parameters are shown in Fig. 7.156A.

The normal cutting speeds for high-speed steels lie within the range of $v_c = 6$ to 45 m/min with feed rates per tooth of $F_z = 0.1$ to 0.4 mm. If bands with inserted cemented carbide teeth are used, the cutting speed can be increased to 200 m/min for steels and up to 2000 m/min for light metals.

In reciprocating sawing (hack sawing), a tool of finite length clamped in a holding frame is used. The feed motion is carried out intermittently only as the tool advances or at a constant perpendicular force.

Circular sawing is sawing with a continuous cutting motion using a circular saw blade. In terms of kinematics and metal cutting technique, circular sawing resembles peripheral milling.

Cutting Tool Materials

Cutting tools consist of the cutting tip, the holder, and the shank. Holders and shanks are designed according to constructional and organizational requirements, such as mating dimensions of the machine, the nature and extent of tool storage and tool changing, or the geometry of the workpiece. The cutting tip is responsible for chip removal. It is subject to mechanical and thermal stresses and chemical attack. The tip consequently wears.

A basic dualism applies to all cutting materials. Harder and more wear-resistant materials are typically less tough and stable under periodically changing loads, as well as under unstable cutting conditions and in intermittent cutting. Materials that cope better with variable mechanical or thermal stresses, are often less resistant to wear. The perfect cutting material combining both features is yet to be developed (Fig. 7.157).

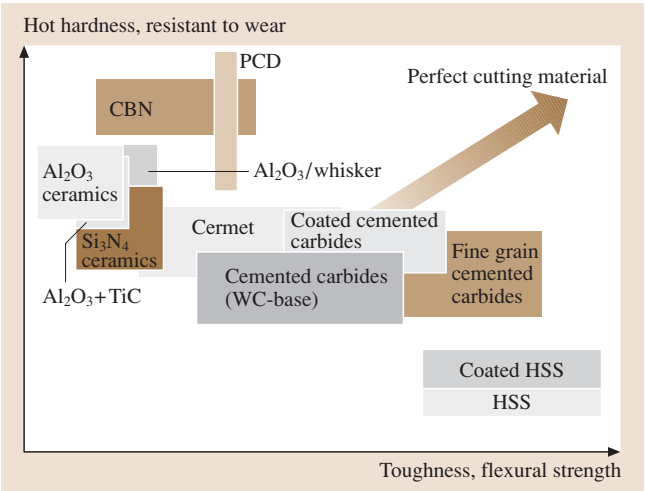


Fig. 7.157 Schematic representation of cutting tool materials (after [7.72])

To overcome this restricting dualism, various cutting materials are manufactured as composite materials. Coating with hard-wearing carbides or oxides produces a separation of functions; the physically (PVD – physical vapor deposition) or chemically (CVD – chemical vapor deposition) vapor-deposited layers provide wear protection and enhanced tribological performance. The tougher substrate performs the supporting function, even under dynamic load conditions. Examples of these coatings include TiN and TiAlN. The latter is applied in dry machining applications due to its higher hardness and enhanced thermal stability compared with TiN [7.74]. Multilayer coatings are also increasingly employed due to their superior tribological performance compared with single layer coatings [7.75].

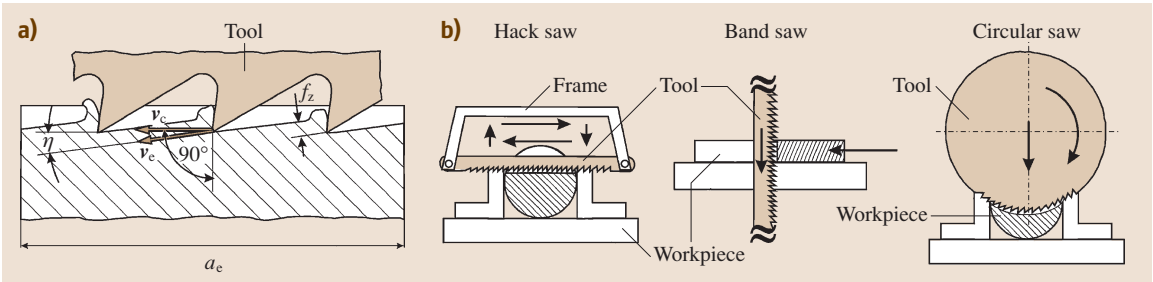


Fig. 7.156 (a) Process kinematics of band sawing and (b) different sawing variants

The primary cutting tool materials include: alloy steels, high-speed steels (HSS), cemented carbides, ceramics and superhard cutting materials (e.g. diamond and cubic boron nitride).

High-Speed Steels. High-speed steels are usually used for drilling, milling, broaching, sawing, and turning tools. Their hot hardness (up to approximately 600 °C) is far superior to that of tool steels (Fig. 7.158). Their hardness results from their basic martensitic structure and from interspersed carbides: tungsten carbides, tungsten-molybdenum carbides, chromium carbides, and vanadium carbides.

The through-hardenability of tools with large cross-sections is increased with molybdenum and/or by alloying with chromium. Tungsten increases the wear resistance and tempered strength, vanadium increases the wear resistance (but is difficult to grind when hard) and cobalt increases the high-temperature hardness. High-speed steels are traditionally manufactured by casting but can also be manufactured by powder metallurgy (PM). PM steels provide enhanced edge strength and cutting edge durability. They are used for thread-cutting and reaming tools. PM high-speed steels with high vanadium contents are easier to grind than their cast high-speed steel counterparts. Their acceptance is only limited because of their higher cost.

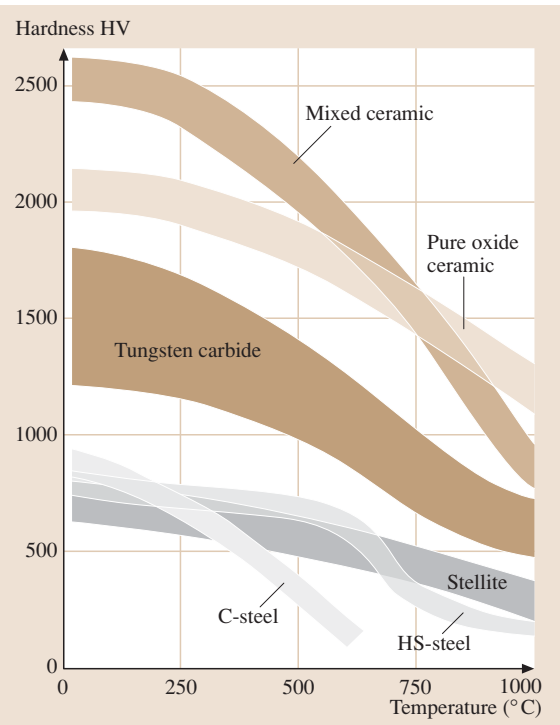


Fig. 7.158 Hot hardness of cutting materials

Coated HSS. High-speed steels are usually coated by physical vapor deposition (PVD), which is reactive ion plating performed at low temperatures in order to stay below the tempering temperature during the process. Simple shapes such as indexable inserts can be treated by chemical vapor deposition (CVD) followed by rehardening. Coated tools (drills, taps, hobs, form turning tools) can have a significantly increased tool life compared to uncoated tools.

A coating of single layers or multilayers deposited on the cutting tools improves its performance. The most frequently used coating materials are titanium carbide (TiC), titanium nitride (TiN), titanium-aluminum nitride (TiAlN), and titanium carbonitride (TiCN).

Coated HSS provides increased hot hardness and wear resistance combined with a high ductility and bending strength of the tool. Tool life can be increased up to 2–2.5 times at increased cutting speed by 20–40%. The most important physical-mechanical properties and parameters obtained by PVD processes are shown in Table 7.41.

The application areas of coatings for cutting tools strongly depend on their properties. TiN has high thermodynamic stability; it is chemically passive to ferrous materials. It combines high hardness and sufficient plasticity. TiN coated HSS tools are mainly used for operations in continuous cutting, e.g. turning, drilling, etc. with the application of cutting fluid.

TiCN provides higher hardness than TiN, and it shows high resistance against oxidation at increased temperatures. TiCN coating is mostly used for milling at increased cutting speeds.

TiAlN has high hardness and increased resistance against oxidation. This material is recommended for high-speed machining combined with flood type application of cutting fluid.

CrN and TiCrN possess high hardness at a rather high level of plasticity. They have high resistance against corrosion and oxidation. HSS tools coated with CrN and TiCrN are recommend for finishing and semifinishing operations at milling with the application of cutting fluid.

Cemented Carbides. Cemented carbides are two-phase or multiphase alloys manufactured by powder metallurgy with a metallic binder. The materials used are

Table 7.41 Properties of some coatings of HSS tools

Properties	Coating materials				
	TiN	TiCN	CrN	TiCrN[a]	TiAlN[a]
Hardness (HV)	2200–2600	3200–3300	2450–2900	3000–3200	3000–3300
Critical load before coating failure (N)	70–80	65–75	40–50	60–70	50–60
Maximum coating thickness (μm)	10	10	50	20	10
Deposition speed (μm/h)	6–8	6–7	2–4	4–5	4–6
Stability against oxidation (°C)[b]	550	550	700	650–700	800
Friction coefficient (against steel 100Cr6)	0.67	–	0.57	–	0.67–0.75

[a] For composition: 50%Ti-50%Cr; 50%Ti-50%Al;
[b] Heating on air during 1 h

tungsten carbide (WC: α-phase), titanium carbide and tantalum carbide (TiC, TaC: γ-phase). The binder is cobalt (Co; β-phase) with a content of 5–15%. Nickel and molybdenum binders (Ni, Mo) are also used in so-called *cermets* (also cemented carbides). A higher β-phase content increases toughness, while a higher α-phase content increases wear resistance and a higher γ-phase content enhances wear resistance at high temperature. Cermets have high edge strength and cutting edge durability. They are suitable for finishing under stable cutting conditions. The manufacturing of cemented carbides by powder metallurgy permits considerable freedom in the choice of constituents (in contrast to casting).

Cemented carbides retain their hardness up to over 1000 °C (Fig. 7.158). They can therefore be used at higher speeds (by a factor of three or more) than high-speed steels. According to standards (DIN 4990/ISO 513) cemented carbides are classified into the metal cutting application groups P (for long-chipping, ductile ferrous materials), K (for short-chipping ferrous materials), M (for ductile cast iron and for ferritic and austenitic steels), N (for nonferrous metals such as aluminum and copper alloys), S (for superalloys and titanium alloys), and H (for hardened materials, such as steels and cast irons). Each group is subdivided due to toughness and wear resistant grades by adding a number. For example, P02 stands for very hard-wearing cemented carbide and P40 stands for tough cemented carbide. The metal cutting application groups do not correspond to grades of cutting tool material, but to the application areas of the finished cutting tools.

Most cemented carbide cutting tools are coated with titanium carbide (TiC), titanium nitride (TiN), aluminum oxide (Al_2O_3), or chemical or physical combinations of these substances. The coatings are applied by CVD or PVD techniques. They are used to achieve longer tool lives or higher cutting speeds. They broaden the range of use of a grade. Coated cemented carbides should not be used for nonferrous metals, high-nickel ferrous materials or, because of the edge rounding caused by the coating process, for precision or ultra-precision machining (cermets are better suited for this purpose). Intermittent cutting and milling requires coatings of especially high bonding strength, which can be influenced by process control during coating.

The application areas of carbides fall into six groups as shown in Table 7.42. The classification is based on the properties of each grade and the machining conditions, type of material being machined, and chip formation.

Ceramics. Ceramic cutting tool materials are single-phase or multiphase sintered hard materials based on metal oxides, carbides, or nitrides. In contrast to cemented carbides, no metallic binders are needed and the material provides high hardness even at temperatures above 1200 °C. Ceramic inserts are generally suitable for machining at high cutting speeds, usually exceeding 500 m/min.

The use of aluminum oxide ceramic is restricted by its lower bending strength and fracture toughness compared with cemented carbide. In intermittent cutting with alternating mechanical and thermal stresses, microcracking, crack growth with microchipping or total fracture can occur. This effect greatly depends on the nature and composition of the ceramic. The change from single-phase materials (Al_2O_3) to multiphase materials has improved toughness considerably: today Al_2O_3 containing 10–15% ZrO_2 (transformation toughening), Al_2O_3 with TiC (dispersion strengthening) or Al_2O_3 reinforced with SiC whiskers (high

Table 7.42 Classification of cemented carbides according to ISO 513:2004-07

Groups of cutting carbide materials			Subgroups of application		Hardness, wear resistance, cutting speed	Toughness, bending strength, feed speed
Designation	**Colour**	**Material machined**				
P	Dark blue	Steel: all steel, cast steels, except stainless steels	P01 P10 P20 P30 P40 P50	P05 P15 P25 P35	↑	↓
M	Yellow	Stainless steels: austenitic stainless, austenitic-ferritic steels, cast steels	M01 M10 M20 M30 M40	M05 M15 M25 M35	↑	↓
K	Red	Cast irons: cast irons with lamellar graphite, cast irons with spherical graphite	K01 K10 K20 K30 K40	K15 K25 K35	↑	↓
N	Green	Nonferrous metals: aluminum and other nonferrous metals, nonmetal containing materials	N01 N10 N20 N30	N05 N15 N25	↑	↓
S	Brown	Special alloys and titanium: heat resistant special alloys, ferrous – nickel-cobalt, titanium and titanium alloys	S01 S10 S20 S30	S05 S15 S25	↑	↓
H	Grey	Hard materials: tempered steel, high-strength pig-iron, chilled cast pig-iron	H01 H10 H20 H30	H05 H15 H25	↑	↓

toughness) are used. The main field of application is the turning of flake graphite cast iron at high cutting speeds, though the turning of steel is also feasible. Additions of up to 40% TiC to the Al_2O_3 ceramic (black mixed ceramic) increase toughness and edge strength. These are used for hard machining and the finishing of cast iron.

Silicon nitride (Si_3N_4) exhibits many excellent cutting material qualities (high strength, hardness, oxidation resistance, thermal conductivity, and resistance to thermal shock) owing to strong covalent bonding of the elements. Here, fracture toughness is not a primary limiting factor in selecting application areas. Sintered Si_3N_4 ($\rho = 3.1\,g/cm^3$, $R_m = 650\,MPa$), hot-pressed Si_3N_4 ($\rho = 3.2\,g/cm^3$, $R_m = 700\,MPa$) and the material system Y-Si-Al-O-N are all used as cutting tool materials. The manufacture and use of Si_3N_4 is currently limited by the sintering auxiliaries (e.g. magnesium oxide, yttrium oxide), which determine the glassy phases in the cutting material. When machining steel or ductile cast iron, failure occurs due to severe

Table 7.43 Properties of oxide ceramics

Properties	Al_2O_3	Al_2O_3-ZrO_2	Al_2O_3-TiC	Al_2O_3 with SiC whiskers
Hardness (30 HV)	2000	2000	2200	2400
Young's modulus (GPa)	390	380	400	390
Bending strength σ_B (MPa)	350	600	600	600–800
Fracture toughness K_{1C} (MPa $m^{1/2}$)	4.5	5.8	5.4	6–8
Coefficient of thermal expansion α (10^{-6} K^{-1})	7.5	7.4	7.0	–
Thermal conductivity λ (W/(m K))	30	28	35	35

wear. Si_3N_4 is suitable for turning and milling of gray cast iron, for highly intermittent cutting actions, and for the turning of high-nickel content materials. The properties of some oxide ceramic materials are given in Table 7.43.

Superhard Cutting Tool Materials. Superhard cutting tool materials include polycrystalline diamond (PCD), polycrystalline cubic boron nitride (PCBN), monocrystalline diamond (MCD), and various forms of chemical vapor deposition diamonds (both thin-film coatings and self-supporting thick-films). The polycrystalline diamond is typically manufactured as a *backed* 0.5–2.0 mm layer of superhard composite on a cemented carbide substrate. PCD is used to machine nonferrous metals, including metal-matrix composites (MMC), wood, composites, stone, and certain cast irons. Due to a well-defined maximum operating temperature of between 700–800 °C, it cannot be used for machining of steels. PCBN is manufactured as either a carbide-backed or solid material. Different grades of PCBN are used for the machining of gray, white, and high-alloy cast irons, and hardened steels. Coatings on PCBN tools are becoming increasingly popular as with cemented carbides. Monocrystalline diamond tools are used for high-precision and ultra-precision machining of aluminum, copper, electroless nickel, glass, plastic, and silicon where, for example, surface finish requirements of several nm and form errors of less than 0.1 µm are common [7.76]. Being a single crystal material, it is possible to produce an extremely sharp cutting edge, within the range of 10–100 nm. Thick-film CVD diamond is generally considered as lying between monocrystalline and polycrystalline diamond, in terms of properties and behavior in application.

In Table 7.44 conditions of cutting parameters and application areas of PCB and PCD tools are shown.

Table 7.44 Application of and cutting parameters for PCB and PCD tools

Work material	Cutting speed (m/min)	
	Turning	Milling
PCB		
Structural and tool steels (without thermal treatment) (< 30 HRC)	–	400–900
Hardened steels (35–55 HRC)	50–200	200–400
Hardened steels (55–70 HRC)	40–120	80–300
Grey iron, high strength cast iron (150–300 HB)	300–1000	600–3000
White cast iron and hardened cast iron (400–650 HB)	40–200	150–800
PCD		
Aluminum and aluminum alloys	600–3000	600–6000
Al-Si-alloys (Si < 20%)	500–1500	500–2500
Copper and copper alloys	300–1000	300–2000
Composed nonmetallic materials and plastics	200–1000	200–2000
Wood	–	2000–4000
WC-Co carbides	15–30	15–45

Tooling

The design of tools has to take all possible impacts into account which can occur during the specific process, comprising mechanical, thermal and chemical loads, as well as flexibility, changeability, maintenance, and costs. Therefore solid tools, tools with brazed tips, and tools with indexable inserts are in use.

In turning the toolholders have to be mounted to the different types of turrets on a lathe, taking into account tool changing time and collision contours. In the case of driven tools like milling shank type cutters have to be clamped in specific toolholders, which are used as interface between the actual cutting tool and the main spindle of the milling machine (Fig. 7.159). The stiffness of the cutters and the toolholders is important for the required surface quality and has a strong influence on the probability of chatter in milling.

In the case of high performance or high speed cutting conditions significant attention has to be paid to the quality of the machine–tool-interface regarding unbalance, stiffness, and damping [7.77, 79]. The machine facing side of these toolholders is standardized. Most often short tapers, steep-angle tapers, or hollow shank tapers are used. The toolholders are fixed in the main spindle by mechanical or hydraulic power.

A number of different toolholder systems are available on the market. This system, the machine tool and the tool changer determine the possible chip-to-chip time in case of a change of the tool [7.80].

Microcutting

Micromanufacturing is defined as the production of components with feature sizes in the range of 1–500 µm. Microcutting is therefore a process used for the production of components of micron dimensions with submicron form accuracy and surface roughness to within a few tens of nanometers [7.81, 82], achieved by mechanical removal of material using defined edge tools. Common microcutting operations include microturning, microdrilling, and micromilling. Examples of features produced by microcutting are shown in Fig. 7.160 while microcutting tools are shown in Fig. 7.161.

The scale of the tool cutting edge geometry, often simplified as a single radius of curvature, is generally similar to the scale of the undeformed chip thickness (as indicated in Fig. 7.162b). The result is a *size-effect* due to a more negative *effective rake angle* as the undeformed chip thickness is reduced. There is also an increase in friction, elastic and ploughing energies, relative to the energy expended in material removal, and therefore an increase in specific cutting

Shank type tool clamping devices	 Precision collet chuck	 Hydraulic expansion-chuck	 Heat shrink chuck
rpm	< 20 000	< 30 000	< 30 000
Torque	Average	High	Very high
Accuracy	< 3 µm	< 3 µm	< 2 µm
Damping	Good	Very good	Very good
Stiffness	High	Very high	Very high
Handling	Average	Very simple	Complex

Fig. 7.159 Tool clamping devices for advanced applications (after [7.77])

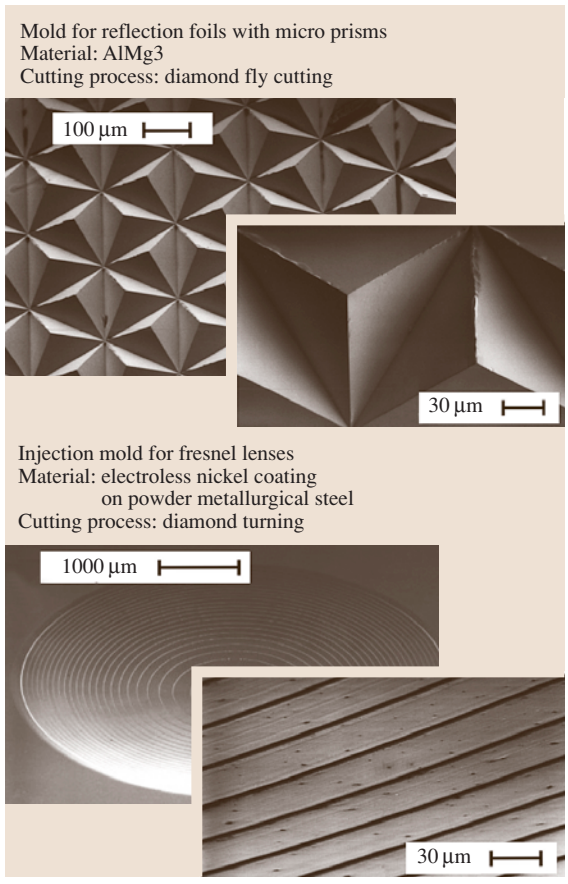


Fig. 7.160 Features produced by diamond microcutting (source: LFM, Bremen) (after [7.78])

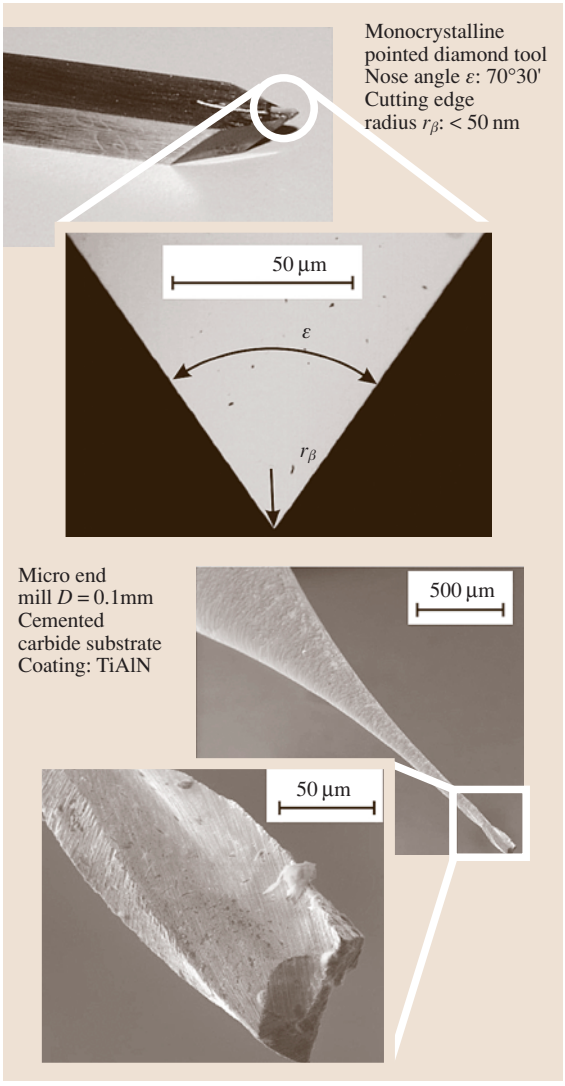


Fig. 7.161 Microcutting tools for turning (source: LFM, Bremen) and milling (source: IPK, Berlin) (after [7.76,84])

energy. The increasing negative rake produces hydrostatic stresses in the near surface enabling ductile mode machining of brittle materials (optical glasses, ceramics, etc.) [7.83]. An isotropic material size-effect also pertains whereby the material strength approaches the theoretical value due to the reduction in dislocation density in the cutting zone as the undeformed chip thickness decreases.

The work material microstructure may also affect the microcutting process performance. For example, the grain size and anisotropy of polycrystalline (PC) metals and alloys can affect the surface finish. Figure 7.162b shows a simplified model of this mechanism. The modulus of elasticity E of a PC material varies in different atomic lattice directions. After the cutting tool has passed, as shown, the grains return to equilibrium positions (the *springback effect*) but to different surface levels due to the differences in the crystal orientations and modules of elasticity.

Tool materials for microcutting must be of high hardness. Diamond tools are generally used for cutting nonferrous metals, ceramics, semiconductors, and glasses. Tungsten carbide is used for ferrous-based metals where diamond tools are limited due to their wear conditions (graphitization, oxidation, diffusion, and carbide formation). Single crystal diamond cutting tools are preferred for microcutting as they can have cutting edge *radii* in the order of 10 nm. With polycrystalline tools, such as tungsten carbide-based tools, the achievable edge sharpness is limited by the grain size. Ferrous alloys have recently been machined with single crystal diamond tools using ultrasonic elliptical vibration cutting to lower cutting temperatures [7.85].

The mechanics of tool and machine tool design must be considered in microcutting. Small diameter tools for microdrilling and micromilling require very high spindle speeds. Furthermore the precision of a machine tool has to be increased in order to maintain a given relative dimensional tolerance. As a component and associated load are reduced in scale, the structural stiffness will decrease, and the natural frequency and internal stresses will increase. Finally, the ratio of total surface area to workpiece (feature) volume is increased in microcomponents, making the surface integrity of the component more significant in comparison with bulk properties.

High-Speed Cutting

The development of high-efficiency machine tools and new cutting tool materials facilitated the application of high-speed cutting (HSC) in metal working.

The main objective of HSC application is to the reduce machining time and increase the removal rate. High cutting speeds lead to lower cutting forces, higher machining accuracy, and improved surface quality compared to traditional cutting [7.86].

The HSM concept was initially developed for the aerospace industry, especially for machining titanium alloys, but was soon adapted by other industrial branches using different materials. At present, a range of materials are being machined with HSM, including nickel-based alloys, titanium, steel, cast iron,

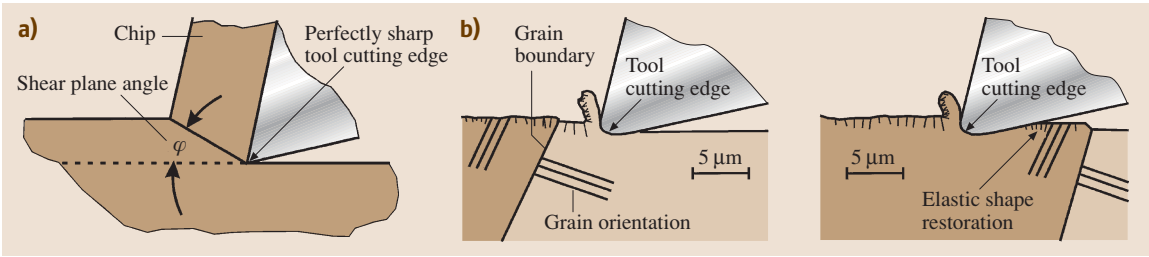


Fig. 7.162 (a) Conventional orthogonal cutting and (b) microcutting showing finite tool cutting edge geometry and workpiece anisotropy

aluminum, and reinforced plastics. The industries using HSM include the aerospace, automotive, and tool/die industries, just to mention the most important users, primarily for manufacturing parts made from aluminum, titanium, and steel alloys. There are a number of extremely important challenges affecting high speed machining, namely the need for hard machining, environmentally friendly hard machining, and the machining of free form (sculptured) surfaces, high material removal rates, increased productivity and quality.

In aerospace the wing spar and stringer components are made from high-grade aerospace aluminum materials. Typically at least 85% of the aluminum billet is turned into chips, which requires excellent off-line work preparation to eliminate machine tool loading time waste, powerful chip-management, and cutting fluid management systems. It is extremely important in the aerospace industry that the HSM technology does not induce residual stresses into the machined component, thus it drastically reduces the problem otherwise caused by fatigue. It is also important that thin walls can be machined easily without deformation.

There are a number of major automotive industries in the world, who also use HSM for making their strategic components in combination with hard cutting to replace the very expensive grinding technology.

In the tool/die industry tools and dies are normally made of difficult-to-machine alloys, which have to be heat treated for high production performance. This represents an extremely high challenge for high speed machining, including machine tools, cutting tools, and technology alike.

In HSM the maximum cutting speed depends upon the workpiece material, the type of cutting operation, and the cutting tool used.

Salomon proposed to use high cutting speeds when machining various materials [7.87]. This idea was based on the hypothesis that when the cutting speed increases to a critical value, the cutting temperature decreases. Some research has been carried out with cutting speeds within the range of $v_c = 47\,000–132\,000\,\mathrm{m/min}$. The most general definition of HSM is the economical utilization of resources and functions to remove the greatest amount of material in the shortest time span by mechanical means.

The achievable maximum cutting speed largely depends on the available high-speed spindle technology and tool technology, including the tool diameter. The high-speed range lies between 800 and 10 000 m/min, depending on the cutting material, productivity, and accuracy [7.88].

In principle, the higher the cutting speed, the more energy is required to remove material [7.89]. The generated heat causes a softening of the material, thus the energy requirement is reduced. During high-speed machining friction and the length of the contact area between the chip and the rake face of the tool reduces contact time. This results in corresponding heat distribution in the cutting system and decreased intensity of heat flow to the tool and workpiece.

The increased productivity of processes using HSC is attributed to the improved material removal rate. Reduction of cutting forces, thermal emission level at the cutting area and thermal-mechanical effect on the machined surface, as well as its elastic form recovery allows considerable improvement of quality and accuracy of machined parts.

HSC has a number of competitive advantages over traditional machining:

- Substantially increased material removal rate
- Excellent surface finish with superb surface integrity
- Low cutting forces, which allow for the machining of thin wall sections
- High frequency system that does not cause any vibration during machining

Table 7.45 Examples of high-speed cutting technologies

Material	Cutting tool	Machining operation	Cutting speed (m/min)
Al-alloys Mg-alloys	Coated carbides, PCD	Milling	1000–7000
High-temperature steel, graphite, copper	Coated carbides, cutting ceramics, PCB	Milling	350–2000
Reinforced plastics, nonferrous metals	Coated cermets	Drilling	100–300

- *Cold* machining, whereby the chips take away most of the heat generated in the process
- Increased accuracy of machining

Table 7.45 shows some high-speed cutting applications.

The efficiency of high-speed cutting becomes apparent when cutting hardened steels with hardness above 45 HRC. HSC of hardened steels and refractory materials is only possible at a small depth of cut and with small wear land of the cutting tool. Machine tools for high-speed cutting should have high rigidity, be equipped with high pressure or minimum quantity lubrication system.

In any case the definition of high cutting speed is material dependent. In Fig. 7.163 an attempt is presented to deduce the area of high cutting speed v_{HSC} from material properties. Extensive research has led to the conclusion the tensile strength of the workpiece material is suitable for a definition of v_{HSC} by means of an empirical equation [7.90].

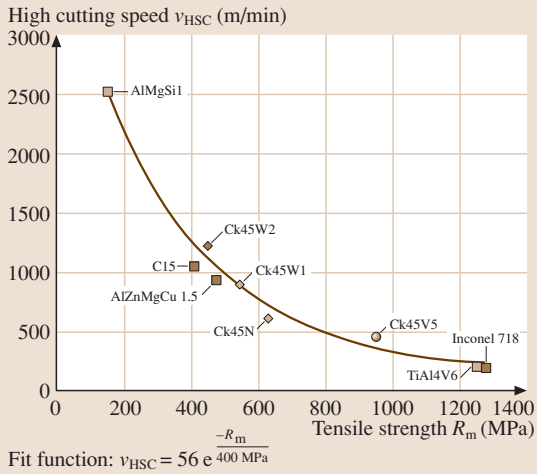


Fig. 7.163 Material depending definition of HSC (after [7.90])

High-Efficiency Machining (HEM). High-efficiency machining means cutting a part in the least amount of time, which is the real target function in machining. High-efficiency machining can be achieved by a variation in the feed rate. Typically, high-speed machining is accomplished with very small axial cut depths in order to achieve good surface finish and avoid damage to the cutter, workpiece, or spindle. Feed rate optimization software can be employed to achieve better cutting efficiency with greater axial depths at high feed rates similar to those of HSM and to protect the cutter in those few places where the chip load momentarily increases. Constant chip load tool paths allow optimum use of the cutter's strength and the machine's speed and power. The software detects conditions under which the chip load would be too high and adjusts the feed rate to a more reasonable level. It returns to the higher feed rate when it is suited [7.91].

Thin-walled components are used widely in the aerospace industry. However, thin-walled components very easily deflect under the cutting forces during the cutting process, which results in compromised machining precision and efficiency. Some solutions have been proposed for the control of machining deflection, such as the NC compensation approach by a tilting tool.

Using this method, the machining deflections are analyzed by FEM. Then the feed rate is adjusted to the degree of deflection by CNC compensation in order to cut off the excessive material due to the deflection. Thus, the thin-walled component can be machined with high precision and high efficiency on a 5-axis CNC machining center.

Hard Machining

Machining of steels of hardness beyond 45 HRC is often required in metal cutting; it is called hard machining. Therefore hard machining, turning, milling, and drilling, are viable alternatives to grinding.

Hard machining is characterized by a low friction coefficient in the cutting area, the application of a rel-

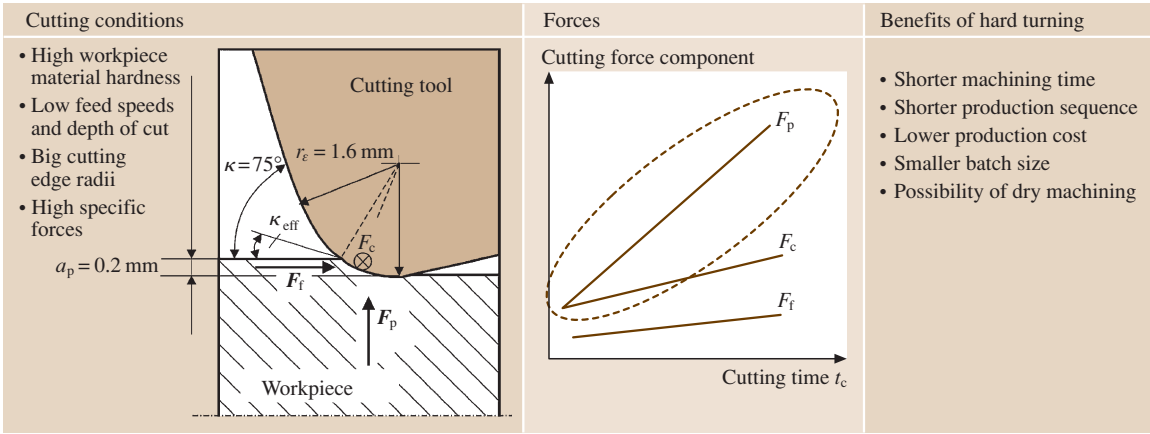


Fig. 7.164 Principle and benefits of hard turning (after [7.92])

atively high cutting speed, low feed and depth of cut, which result in reduced chip–tool contact length. Due to the redistribution of the generated heat which causes a softening of the material and to the high compression load on the workpiece material, machining efficiency increases. The major breakthrough of this technology was connected with the introduction of suitable cutting tool materials as ceramics and CBN for turning and coated tungsten carbides for milling and drilling. These developments made hard machining a serious competitor for grinding and electro discharge machining. As an example of successful application of hard machining the conditions and benefits of hard turning are presented in Fig. 7.164. Important to mention is the fast rising and dominating passive or back force F_p, which needs an excellent lathe with superior stiffness to compete with grinding. Also the possibility to apply dry machining is a major benefit, allowing all costs and problems related to cutting fluids to be avoided.

7.3.2 Machining with Geometrically Nondefined Tool Edges

Fundamentals

Machining with geometrically nondefined tool edges is cutting by the mechanical action of cutting edges on the material. The cutting edges are formed by grains of hard material randomly shaped and arranged. They are either bonded together into a tool (grinding, honing) or are used loose (lapping, abrasive blasting). The cutting edge geometry is not described with reference to a single grain. The individual cutting edge is geometrically nondefined. The processes are divided into the following subgroups:

- Grinding with rotating tools
- Belt grinding
- Honing
- Lapping
- Free abrasive grinding/abrasive tumbling
- Machining by abrasive blasting

The common factor in these processes is that the grains of hard material generally form several cutting edges. The important cutting edge angles for chip formation, the clearance angle α, the rake angle γ, and the wedge angle β are only indicated by means of statistical parameters such as mean values or distributions. On average, sharply negative rake angles and large contact and friction zones are formed between the grain and the workpiece. The cutting edges penetrate only a few micrometers into the material. The undeformed chip thickness distribution depends on the positioning of the cutting edges in the mixture of grains (microtopography of the cutting zone) and the geometry of the machined workpiece surface. Not only chip removal but also elastic and plastic deformations without chip removal take place.

High normal forces result between tool and workpiece at the predominantly negative rake angles of the cutting edges. They lead to elastic deformations in the machine (stretching of the frame and deflection of the spindle), the tool, and the workpiece. The amplitude of deformations may be comparable to the normal small feed motions. Therefore a distinction should be made between the theoretical and the actual feed motion (Fig. 7.165).

Machining processes with geometrically nondefined tool edges are frequently used as final machining

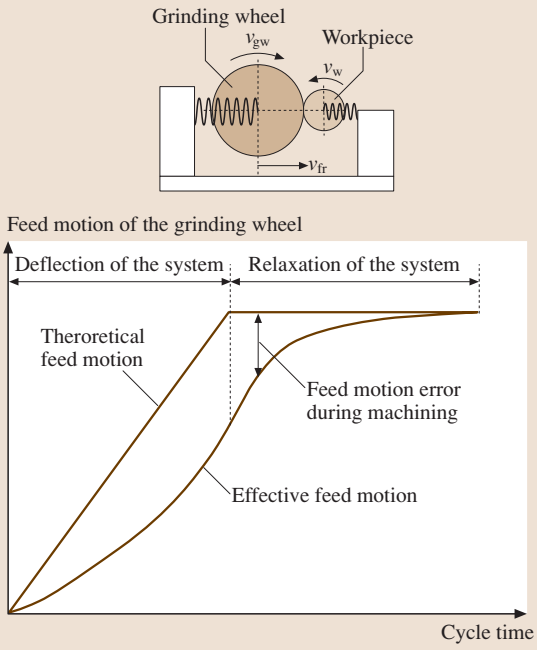


Fig. 7.165 Feed motion errors in grinding due to elastic deformations in the machine tool–workpiece-system

processes for workpieces subject to exacting quality requirements. Figure 7.166 shows a comparison of various precision machining processes with regard to operational results and efficiency. It can be seen that the grinding processes achieve high rates of material removal, while honing and lapping are able to produce the best surface qualities.

The chip formation mechanism when using geometrically nondefined tool edges differs from that of machining with geometrically well-defined tool edges (Fig. 7.167). In phase 1, the often sharply negative rake angle of the individual grains causes elastic deformation of the material. In phase 2 plastic deformation occurs, while in phase 3 the actual chip removal takes place. A large amount of friction occurs between the individual grains and the workpiece.

The mechanical energy supplied is almost exclusively converted into heat. Figure 7.168 shows a qualitative distribution of the heat flows around an individual grain. Most of the heat generated flows into the workpiece, while a small part finds its way into the grain, the bond and the surroundings (cooling lubricant, atmosphere). Temperature increases in the workpiece may harm its surface integrity. This is manifested in thermally induced residual stresses, structure changes or cracks, which influence the subsequent behavior of the workpiece in service. The use of abrasive materials (CBN, diamond) and bonds with superior thermal conductivity reduces the proportion of heat that penetrates into the workpiece [7.93, 94].

In machining with geometrically nondefined tool edges, the use of cooling lubricants is important for the end result. The cooling and lubricating effect can reduce tool wear. Furthermore, the temperature of the workpiece is reduced and the danger of thermal damage to its surface is decreased. The lubricants used are nonwater-miscible (oils) and water-miscible (emulsions, solutions) cooling lubricants, the effect of which can be further improved with additives (polar and extreme pressure (EP) additives to improve the lubri-

	Grinding	Honing	Lapping
Accuracy of shape	+	0	++
Dimensional accuracy	+	++	++
Positional accuracy	+	–	–
Surface quality	0	+	++
Positive influence on surface integrity	0	+	+
Material removal rate	++	0	–
Manufacturing costs	–	+	++

– minor influence + major influence
0 moderate influence ++ superior influence

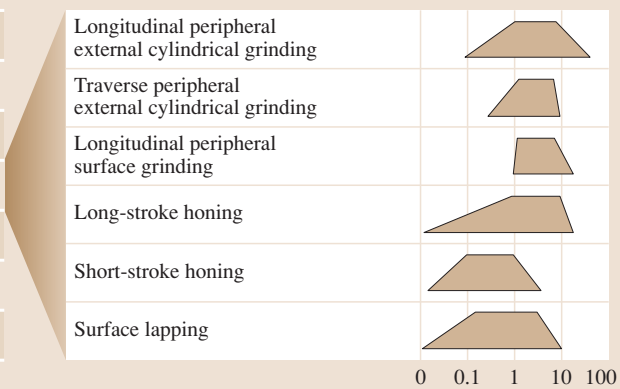


Fig. 7.166 Economic and technological comparison of various precision machining processes

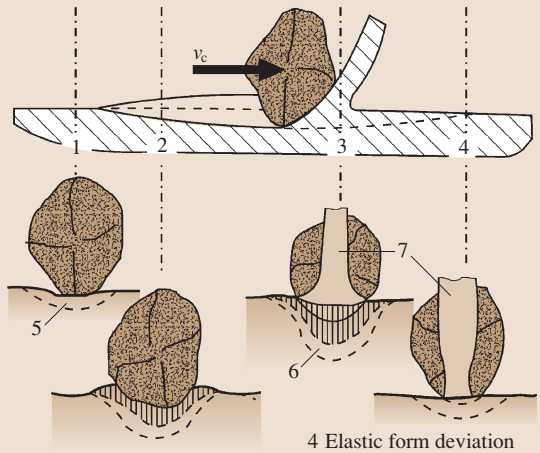


Fig. 7.167 Phases of chip formation during grinding

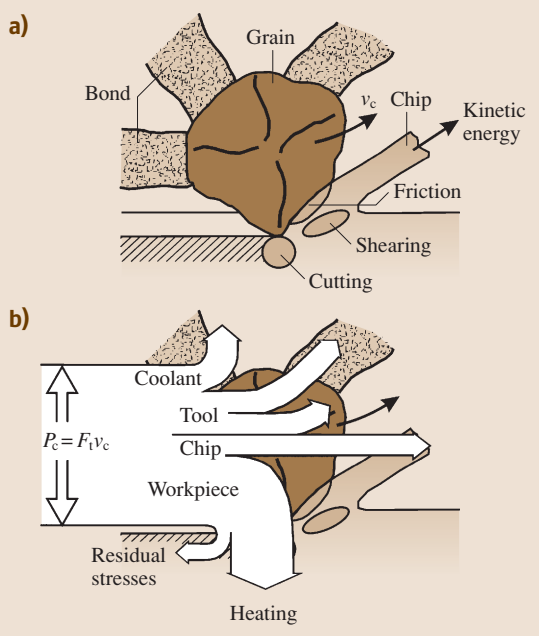


Fig. 7.168a,b Energy conversion: (a) Effects of energy conversion; (b) energy flows

cating effect, antifoaming agents, biocides, and rust inhibitors). The cooling effect depends on physical parameters: specific heat capacity c in kJ/(kg K), heat transmission coefficient α in W/(m^{2}K), thermal conductivity λ in W/(m K), heat of evaporation I_d in kJ/kg, and surface tension σ in N/m. The lubricating effect is described by the tribological parameters of the cooling lubricant.

A comprehensive study on the main effects of coolant type, composition and application in grinding is presented in [7.95]. The correct selection and use of grinding fluids normally results in enhanced process performance, better workpiece quality, and longer tool life. Lower volumetric wear of the wheel and improved surface finishing are achieved when neat oil is used instead of water-based fluids. On the other hand, water-based fluids can reduce workpiece temperature, provide less risk of fire and are generally better suited for the working environment and the machine tool operator.

Grinding with Rotating Tools

Processes. Grinding can be performed in a large variety of motions. The six main classifications are related to the shape of the surfaces produced, namely surface, peripheral, thread, gear, profile, and form grinding. Figure 7.169 shows examples of various motion classifications and tool shapes.

Chip Formation. Material is removed by the penetration of abrasive grains into the workpiece material along a defined path. Owing to the generally unfavorable shape of the cutting edge, the actual chip formation is accompanied by friction and displacement processes. The process is evaluated by calculating statistical averages. Figure 7.170 shows in simplified terms how a comma-shaped chip is formed by the successive action of two cutting edges. While grain 1 has traveled the path AB, the center of the grinding wheel has moved from P_0 to P_1. The next grain 2 will travel the path CD. In this process, the thickness of an average chip increases from 0 up to h_{max}. A simple relationship for the average undeformed chip thickness $\bar{h}$ is obtained by applying the continuity relationship

$$v_{ft} a_e a_p = v_c C V_{sp} a_p \,, \tag{7.172}$$

$$\bar{h} = \frac{v_{ft}}{v_c} \frac{1}{\bar{b} C} \sqrt{\frac{a_e}{d_{eq}}} \,, \tag{7.173}$$

with $\bar{l} = \sqrt{a_e d_{eq}}$, $V_{sp} = \bar{l}\,\bar{b}\bar{h}$ and $d_{eq} = \frac{d_w d_s}{d_w \pm d_s}$ (+ external cylindrical grinding, − internal cylindrical grinding) or

$$\bar{h} = \sqrt{\frac{v_{ft}}{v_c} \frac{1}{rC} \sqrt{\frac{a_e}{d_{eq}}}} \quad \text{with} \quad r = \frac{\bar{b}}{\bar{h}} \,. \tag{7.174}$$

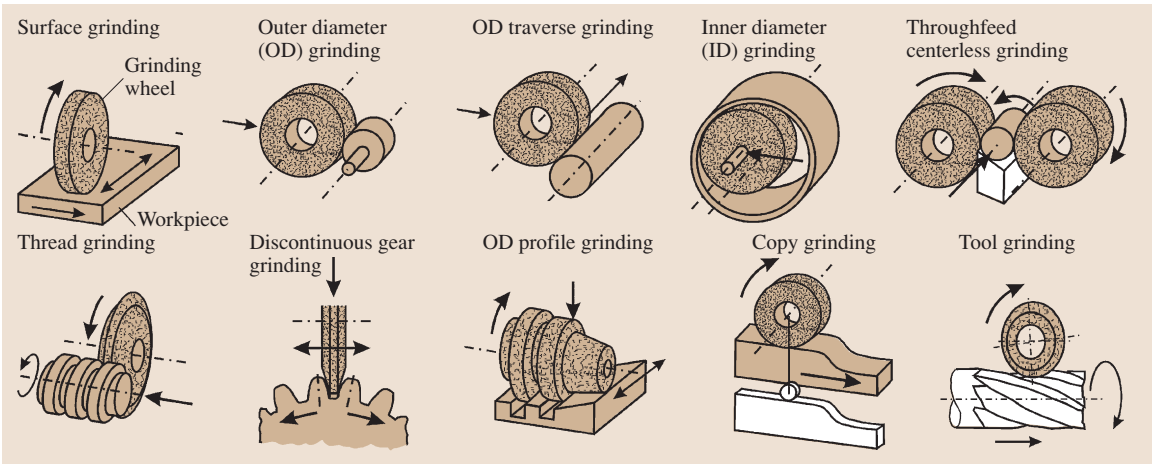


Fig. 7.169 Various grinding processes

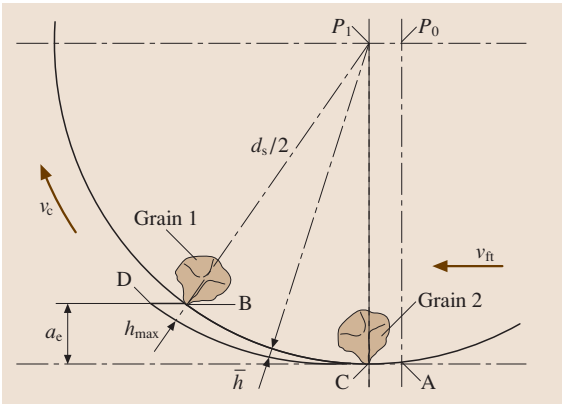


Fig. 7.170 Schematic chip formation in surface grinding

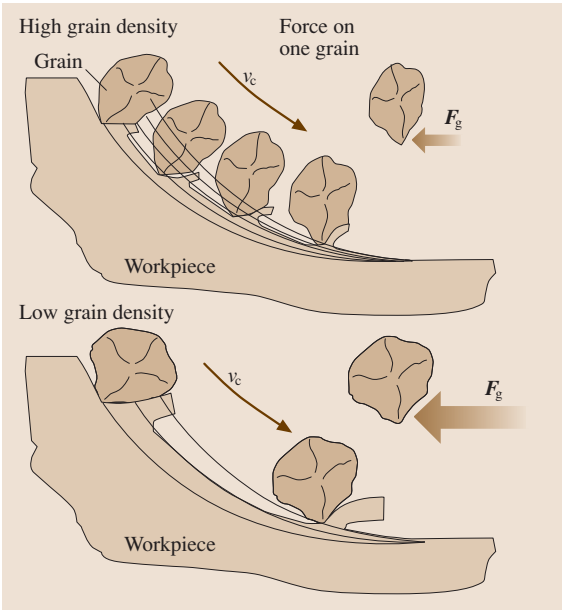


Fig. 7.171 Influence of grain concentration on the chip formation mechanism

The symbols are as follows: $\bar{h}$ is the average (undeformed) chip thickness, $\bar{l}$ is the average (undeformed) chip length, $\bar{b}$ is the average (undeformed) chip width, v_{ft} is the feed speed of the workpiece, v_c is the cutting speed, a_e is the depth of cut a_p is the width of engagement (width of grinding), d_{eq} is the equivalent grinding wheel diameter, d_s is the grinding wheel diameter, d_w is the workpiece diameter ($\rightarrow \infty$ in surface grinding), C is the number of active cutting edges per unit of surface area of the grinding wheel, and r is the ratio of average chip thickness to average chip width. The maximum chip thickness h_{max} is twice the average chip thickness $\bar{h}$.

The chip thickness is highly dependent on the concentration of abrasives on the grinding wheel surface C (Fig. 7.171). Higher grain density leads to thinner chips and consequently lower forces per grain. The grain density decreases with increasing grain size. The correction of these parameters can influence the wear mechanisms on a grinding wheel. Owing to technical difficulties in measuring the number and distribution of grains, the equivalent chip thickness h_{eq} is often used as a parameter for evaluating the grinding process [7.96]. This parameter is directly related to the grinding forces and

surface finishing

$$h_{eq} = a_e \left(\frac{v_{ft}}{v_c} \right) . \qquad (7.175)$$

Composition of Grinding Wheels. A grinding wheel consists of grains, bonding material (bond), and pores. The abrasive materials used are hard-brittle ones such as zirconium-corundum (ZrO_2 with Al_2O_3), corundum (Al_2O_3), silicon carbide (SiC), cubic boron nitride (CBN), and diamond (C); their hardness values are presented in Fig. 7.172. However, diamond is unsuitable for machining steel, as there is a high chemical affinity between diamond with its carbon-based cubic lattice structure and iron, which leads to rapid tool wear.

Sorting of the grains according to size is accomplished by screening. The basis of all standards is the mesh of the screens through which the abrasive grains pass. The average grain size is determined by the shape (angularity) of the individual grains. Below a certain grain size, sorting can be carried out by settlement out of a slurried suspension of water and grains. For conventional abrasives (e.g. SiC, Al_2O_3) the grain size is usually given in a mesh number; thus high numbers represent small grain sizes in the grinding wheel specification. The grain size for superabrasive wheels (only CBN and diamond) follows the FEPA specification related to the grit diameter.

The bonding material (bond) is chosen according to the requirements of the machining process and of the grain material. Inorganic (ceramic, silicate, magnesite), organic (rubber, synthetic resin, glue) and metallic bonds (bronze, steel, cemented carbide) are used; the most popular bonds are those made of ceramic or synthetic resin. In manufacturing a tool, its structure can be influenced to a certain extent by varying the proportions of grains, bond, and pore volume [7.97].

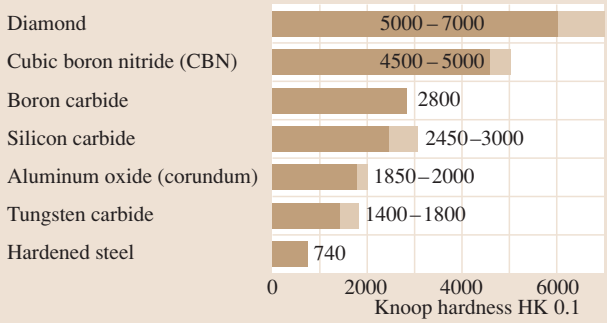


Fig. 7.172 Knoop hardness of various materials

Superabrasive wheels exhibit a particular construction. They consist of a core to which the grinding layer is applied. Usual layer thicknesses are 2–5 mm for resin or vitrified bonds. The core material can be made of metals or fibre reinforced plastics. Superabrasive wheels can grind at higher speeds due to the resistance of the core and the higher hardness of the grains. In the grinding wheel specification for superabrasives the average grain size is usually directly given in µm.

Electroplated bonds are mostly used in superabrasive wheels. They are very thin since they hold only one

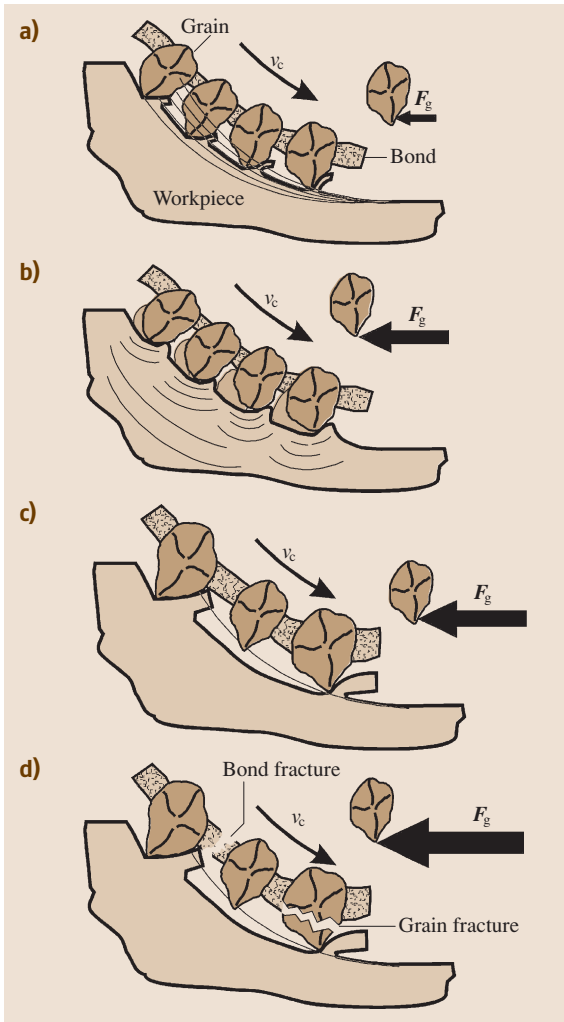


Fig. 7.173a–d Grinding wheel wear mechanisms: **(a)** Sharp wheel with low load per grain, **(b)** dulling and loading, **(c)** sharp wheel with high load per grain, **(d)** grain and bond fracture due to very high load

layer of abrasives. This kind of bond is used for grinding at high speeds due to its high wear and stress resistance. Electroplated wheels are recommended for rough and semifinish grinding. The final surface finish can be obtained with vitrified or resin bonds. The specification of a grinding wheel is standardized in DIN 69100 (for conventional wheels) and ISO 6104 (for CBN and diamond wheels).

Tool Wear. Grinding wheel wear may take place in the grains and the bond. The wear phenomena can be related to dulling or grain/bond breakage (Fig. 7.173). In dulling, wear flat areas are formed on the edges. They lead to increasing grinding forces and temperature. This normally happens when the abrasives are submitted to gentle grinding conditions. When the abrasive grains are under higher loads the grain/bond breakage mechanisms will lead to higher volumetric wear resulting in constant grinding forces and temperature. Surface finish quality may deteriorate and form errors may increase.

Grinding Wheel Conditioning. The purpose of conditioning is to give the grinding wheel the required profile and concentricity (profiling or truing), to produce the necessary grinding wheel topography with sharp grains (sharpening), and to clean the grinding wheel surface [7.98]. The first two operations summed up under the term dressing and are generally performed simultaneously by passing a dressing tool over the surface of the grinding wheel (Fig. 7.174). In some applications a separate sharpening step is necessary, which can be performed preprocess or in-process. It is often performed with special equipment with electrolyte supply (ELID = electrolytic in-process dressing) [7.99] or with contact erosion [7.100]. The main components of dressing tools are metallic bonds mixed with diamond particles; however, there are also single point diamonds, diamond-free steel and ceramic tools. They reach the end of their life when the diamond layer or the tool geometry is worn out.

A special variant of a rotating diamond form roll is called grinding with continuous dressing (CD grinding) (Fig. 7.175). Here, the dressing tool is in engagement during grinding and is continuously advanced radially to the grinding wheel. As a constant grinding wheel profile and a uniform grinding wheel topography with sharp cutting edges are permanently ensured, the material removal rate can be increased considerably [7.101]. With the aid of the machine control system, the dressing tool and the grinding wheel have to be advanced in relation to the workpiece in such a way that the decrease in the diameter of the grinding wheel is compensated. This technique is often used for pro-

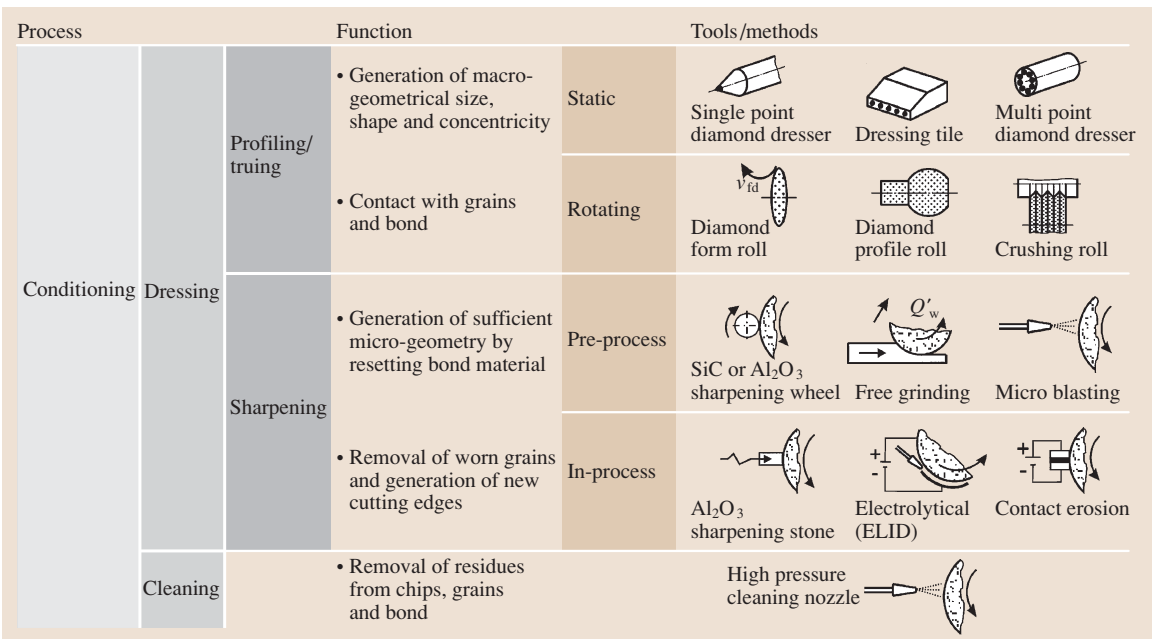


Fig. 7.174 Classification system for grinding wheel conditioning

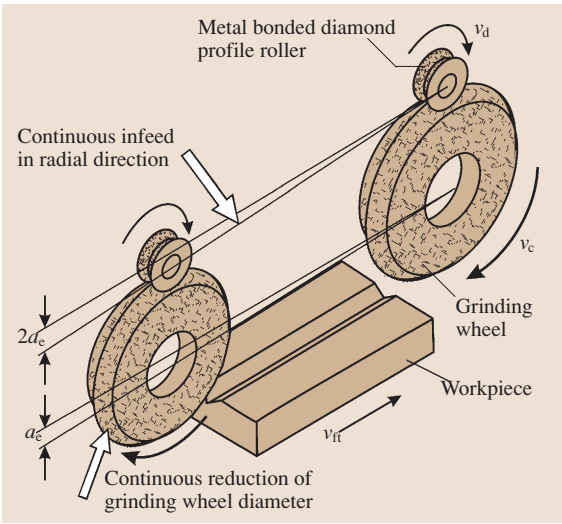


Fig. 7.175 Principle of grinding with continuous dressing (CD grinding)

file grinding of difficult to machine materials like the nickel-based alloys used for turbine blades.

Limits of the Grinding Process. Restrictions on the process arise if the original data, e.g. dimensional accuracy, accuracy of shape, surface quality or condition of workpiece surface integrity, do not lie within the required limits. The interaction of the various influencing variables, such as workpiece, machine setting data, tool, cutting fluid, etc., may be extremely diverse. Mechanical and/or thermal overstressing of the material in the grinding process may adversely affect the characteristics of a ground component [7.102]. Typical grinding defects due to poor process control are visible marks on the surface as a result of regenerative chatter [7.103], hardness and structure changes, tensile residual stresses, and probably even cracks in the workpiece [7.104] (Fig. 7.176). The latter defects can be summarized by the term grinding burn.

Development Trends in Grinding. Grinding has been continuously developed from a traditional precision machining process to improve dimensions, shape and surface quality into a very versatile and efficient manufacturing process. Creep feed grinding, high-speed grinding, grinding with continuous dressing (CD grinding), the growing use of superhard abrasives combined with CNC and sensor technologies have been extensively applied in industry [7.105–107]. With the development of advanced open CNC systems, the control of grinding processes is nowadays including intelligent routines. Sensors allow the detection of contact between grinding wheel and workpiece, collisions and wheel topographic conditions where necessary conditioning/dressing can be automatically activated [7.108]. Gear and other complex profile grinding processes have also been improved with the help of high performance interpolation routines able to increase productivity and quality with high process flexibility [7.109]. Even the fact that grinding might generate a high amount of heat is turned into a benefit by us-

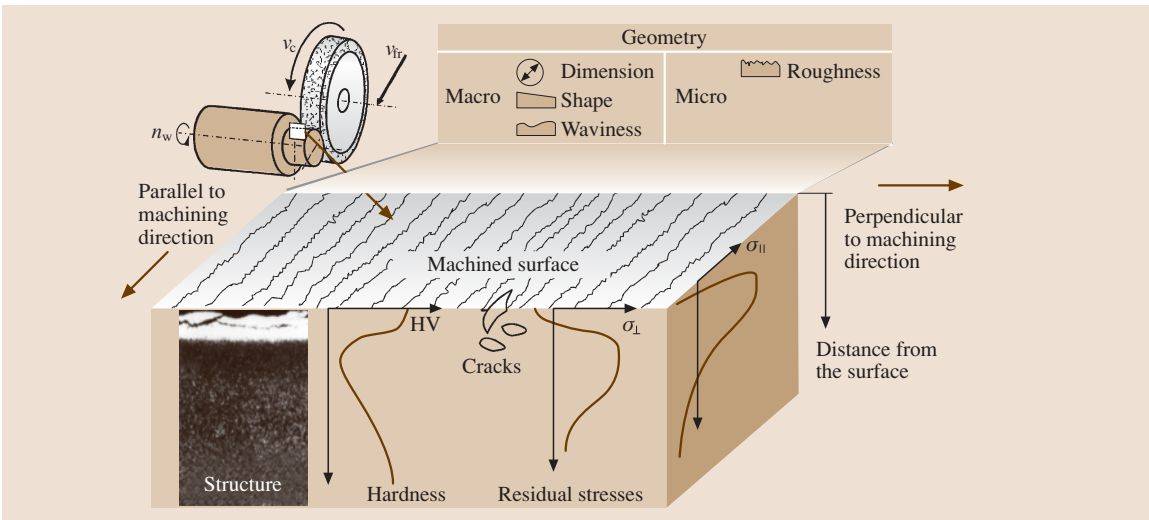


Fig. 7.176 Geometrical and physical quality characteristics after grinding

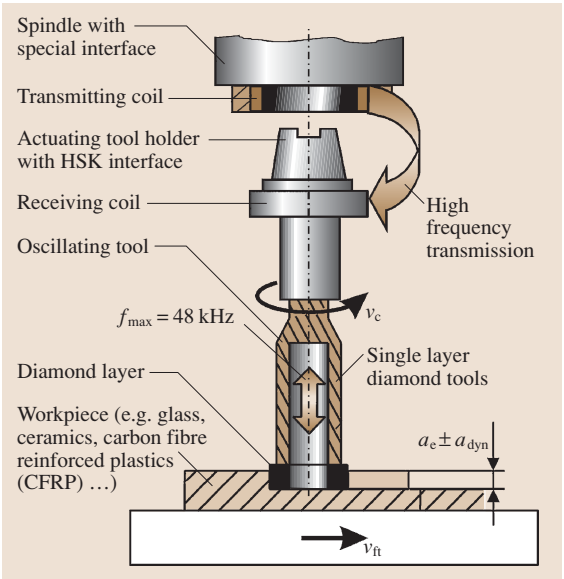


Fig. 7.177 Ultrasonically assisted grinding

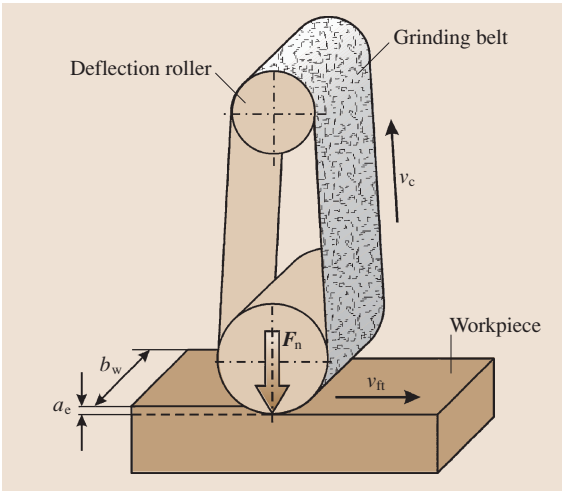


Fig. 7.178 Schematic representation of surface belt grinding

ing this energy for process integrated hardening while grinding [7.110].

Ultrasonic (US) assisted grinding has reached industrial maturity especially for brittle materials like glass or ceramics. While in the first applications the US-vibration was generated in a complex spindle system or from the workpiece clamping side, in latest developments the US-excitation is realized with wireless induction transmission in the tool-spindle interface, allowing the tools to be mounted to the actuating tool holder (Fig. 7.177) [7.111].

Belt Grinding

Belt grinding is grinding with bonded abrasives on a supporting flexible bed. Belt grinding can also be classified according to the surfaces machined, e.g. surface, cylindrical, profile or form belt grinding. Surface belt grinding predominates in industrial applications (Fig. 7.178). Belt grinding is normally performed at a constant perpendicular force F_n (pressing force). Thus, consistent surface qualities can be produced. The material removal rate is determined by the sharpness of the belt (of the active cutting edges). In belt grinding with constant working engagement a_e a constant material removal rate is achieved. Surface quality depends on the condition of the cutting edges. This process is especially suitable for removing large volumes of material with high efficiency [7.112].

The adjustment of the process variables during belt grinding derives from the changes of the cutting edges during working. In contrast to grinding wheels, grinding belts generally consist of a single layer of abrasives. Therefore, the cutting edge zone changes over the period of use, owing to progressive abrasive and bond wear.

Grinding belts consist of four elements: backing (paper, fabric), ground bonding and top bonding materials (phenol resins), and abrasive grains (corundum, zirconium-corundum, silicon carbide). Scattering of the grains on the first coat of bonding material is carried out in an electrostatic field. This ensures that the abrasive grains are aligned perpendicularly (Fig. 7.179). An even grain distribution in belt manufacture compared with conventional gravity scattering can be achieved [7.113].

For higher material removal rates multilayer grinding belts are available, where several abrasive grains are held in a bond and mounted to the backing.

The grinding belt is supported in the working zone by contact elements. A contact disc is used in peripheral grinding, while a contact shoe or beam is used in side grinding. Hard contact rolls made of aluminum or steel are particularly suitable for roughing with coarse grinding belts because they transmit the relatively high grinding forces. Soft, rubber-coated contact rolls are used in finishing with fine grinding belts [7.114]. They absorb the shocks occurring during the process. Conventional applications of belt grinding are grinding of individual sheets and sheet coils, deflashing and deburring, and grinding down of excess metal in welded

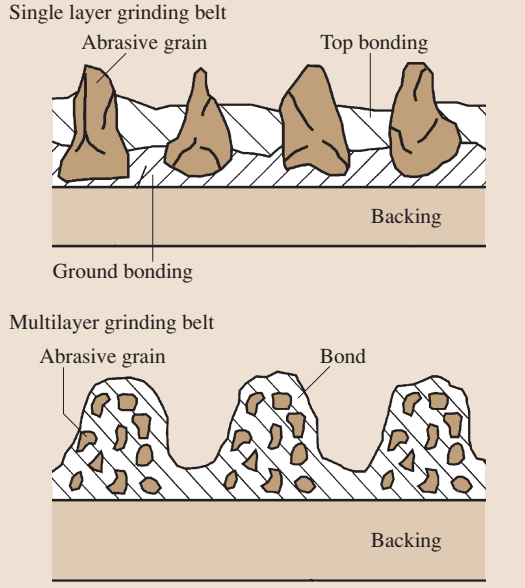


Fig. 7.179 Schematic setup of grinding belts

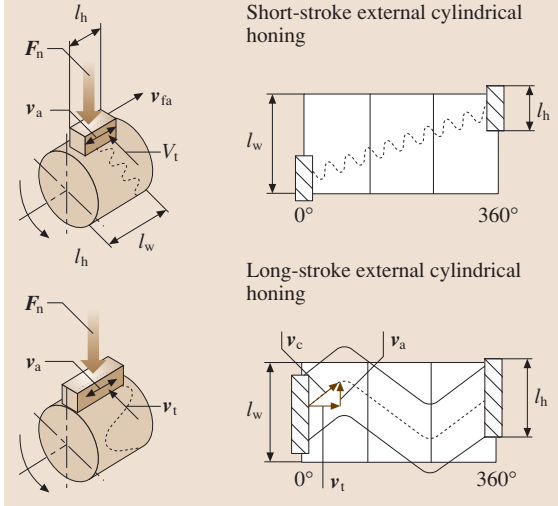


Fig. 7.180 Geometry and kinematics in short- and long-stroke superfinishing

joints. Heavy-duty belt grinding might even work as a substitute for turning and milling of components made of gray cast iron or aluminum alloys [7.112, 113].

Honing, Superfinishing

Honing is performed with a multiple-point cutting tool consisting of bonded abrasive grains using a cutting motion with two components of which at least one is oscillating. The most common honing processes are external cylindrical honing, internal cylindrical honing, and surface honing. The external honing variant is often called superfinishing. According to the oscillation amplitude, a further distinction can be made between two main groups: long-stroke honing and short-stroke honing (Fig. 7.180) [7.115].

Long-stroke honing employs large oscillation amplitudes at a low frequency; in short-stroke honing the oscillating motion is performed at low amplitudes and a correspondingly high frequency. The path curves in Fig. 7.180 depict the motion of a honing strip over a developed workpiece surface.

Owing to the superimposed motion during honing, the workpiece surface exhibits intersecting tracks of the cutting grains, the two tracks enclosing an overlap angle α (Fig. 7.181). The magnitude of the overlap angle α is determined by the selection of the ratio of the axial (v_a) and tangential (v_t) cutting speed components. For workpieces without longitudinal and transverse grooves, the angle α is generally 45°. The cutting speed v_c can be calculated by means of the aforementioned speed components according to $v_c = \left(v_a^2 + v_t^2\right)^{1/2}$. Usually the cutting speed does not exceed $v_c = 1.5\,\mathrm{m/s}$ [7.115, 116].

During the cutting motion the honing stones are pressed against the workpiece surface that is to be machined with a perpendicular honing force F_n, which may be generated by means of various feed systems (Fig. 7.182). In force-dependent feeding, a defined hydraulic pressure p_{oil} is set on the machine. The resulting

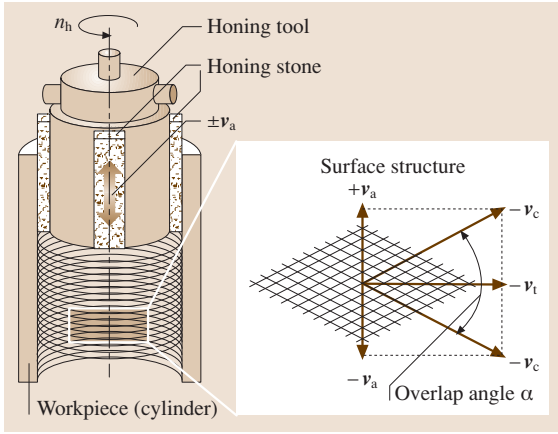


Fig. 7.181 Principle of internal long-stroke honing and resulting surface structure

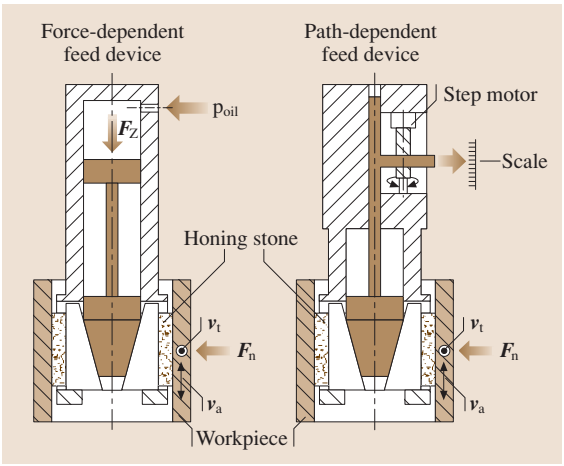


Fig. 7.182 Force- and path-dependent feed devices for honing

advancing force F_z is transmitted to the honing stones via an advancing pin and cones. In path-dependent feeding defined feed paths are generated e.g. with a stepping motor, which produces the perpendicular force F_n on the honing stones.

Important variables influencing the result of honing are the type of abrasive, grain size, type of bond, hardness and impregnation of the honing stones. The types of abrasive can be classified into conventional abrasive materials, such as corundum and silicon carbide and the superhard abrasive materials diamond and cubic boron nitride (CBN). The grain size influences the material removal rate and the surface quality. The achievable surface roughness values are $R_z = 1\,\mu m$ for long-stroke honing and $R_z = 0.1\,\mu m$ for short-stroke honing. Dimensional accuracies and accuracies of shape of the machined workpieces of 1–3 µm are achieved [7.117]. Unlike in grinding, the grains bonded in the honing stone are stressed on more than one axis owing to the oscillating motion. Honing tools are, therefore, self-sharpening.

Cutting fluids are used in honing as well as in grinding. Due to the low cutting speed, however, heating is minimal, so the cooling effect plays a minor role. But the surface contact between the honing stone and the workpiece requires instead a friction-reducing lubrication. For this purpose, pure oil, with additives if required, is generally used.

The applications of honing can also be categorized according to whether long-stroke or short-stroke honing is used. Long-stroke honing is generally used for bores, e.g. cylinders in internal combustion engines. Short-stroke honing is mainly used for machining small cylindrical components, e.g. running surfaces of inner and outer rings and rollers of rolling-contact bearings. Furthermore, the application of gear honing should be mentioned. With the coupled motion of a gear shaped abrasive tool and the workpiece counterpart a high surface quality for gears can be achieved, thus qualifying honing as an attractive gear finishing technique [7.118].

Other Processes: Inside Diameter Cut-Off Grinding, Lapping, and Abrasive Waterjet Cutting

Inside-Diameter Cut-Off Grinding. Inside-diameter (ID) cut-off grinding is a high-precision finish-machining process for hard-brittle materials. It is used to cut rod-shaped materials into thin slices (Fig. 7.183). Besides its applications to optical materials (glasses, glass ceramics), magnetic materials (samarium-cobalt, neodymium-iron-boron), and ceramics and crystals for solid-state-type lasers, this process is also used for semiconductor materials. Single crystal silicon rods are cut into thin slices called wafers.

Compared with conventional abrasive cutting processes, the loss of material in the cutting gap can be reduced by approximately 80% by means of a narrow width of cut. This is a decisive advantage, especially for expensive, high-grade materials.

To achieve the narrow widths of cut that are typical of this process, a comparatively unconventional tool is used for ID cut-off grinding. The basic body of the tool

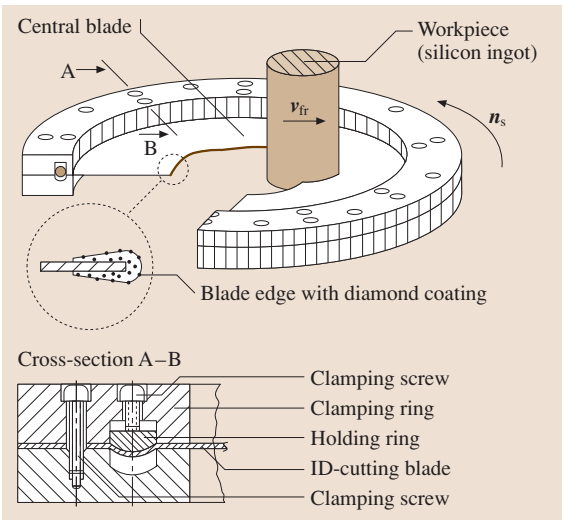


Fig. 7.183 Principle of ID cut-off grinding

consists of a cold-rolled ring of high-tensile stainless steel with a thickness of 100–170 µm.

The inside edge of the cutting blade is coated by electro deposition with a diamond layer in a nickel-base bond; this layer forms the droplet-shaped cutting edge. Grain sizes commonly range from 45 to 130 µm. The width of the cut extends between 0.25 and 0.7 mm. Natural diamond is generally used as cutting material. CBN may also be used for special applications. Workpieces with a diameter of up to 200 mm can be cut. Due to this diameter limitation this process has lost its pole position for wafer manufacturing because of the constantly rising diameters of silicon ingots for wafer production in the semiconductor industry (Fig. 7.185).

To achieve the stiffness at the cutting edge that is necessary for cutting, the cutting blade, which can be compared to an eardrum, is clamped at its outside edge with a special clamping device. This widens the ID cutting blade radially until the tangential stresses at the inside edge reach around 1800 N/mm^2. In cut-off grinding, the workpiece performs a radial feed motion relative to the rotating tool [7.119, 120].

Lapping. Lapping is defined in a German standard (DIN 8589) as metal cutting with loose grains of abrasive material distributed in a paste or a liquid, the lapping mixture, which is carried on a generally shape-imparting counterpart (lap), and the individual grains following cutting paths, which are random. Lapping processes are divided into surface lapping, cylindrical lapping, hole lapping, and ultrasonically assisted lapping (Fig. 7.184).

Surface lapping or plane parallel lapping is carried out on single-disc or twin-disc lapping machines. The lapping discs act as the holder for the lapping abrasive. They are frequently made of perlitic cast materials or hardened steel alloys. The workpieces are usually driven in specific carrier discs, for larger diameter lapping discs often use the planetary gear principle to achieve cycloid motion. A special variant of plane parallel lapping is based on the use of special slurry as liquid to carry the abrasives and to improve the material removal process by chemical activation. This process is called chemical-mechanical planarization (CMP) and is widely used in the semiconductor industry to machine wafers made of silicon or other suitable materials [7.121].

The abrasive consists of lapping powder and fluid in a ratio of 1 : 2 to 1 : 6. Particles of silicon carbide, corundum, boron carbide or diamond are used as lapping powder. The type of grain for a particular application is determined by the material to be machined. Grain sizes ranging between 5 and 40 µm are generally used. Besides high-viscosity oils or similar fluids, water with suitable additives has been more and more frequently used as a transport medium for the abrasive in recent years. The main purpose of lapping fluids is to cool the workpiece and ensure chip removal from the effective zone.

Lapping is a precision or ultra-precision machining process for the production of functional surfaces of optimum surface quality. Surface roughness values down to $R_t = 0.03$ µm, flatness, and plane parallelisms of 0.2 µm are achieved. Besides the semiconductor application other typical fields of lapping are the machining of precision cemented carbide tools, calliper gauges, or hydraulic rams.

A special category is ultrasonically (US) assisted lapping, which is particularly suitable for machining hard-brittle materials, e.g. fully sintered ceramic components [7.122, 123]. The process is carried out by moving a complex shaped tool called sonotrode downwards into the workpiece while hammering the loose abrasive particles, which are supplied via a liquid to the contact zone, into the workpiece surface for material removal. If rotational symmetric shapes are to be produced, nowadays US-grinding is substituting for this lapping process due to its higher efficiency (Fig. 7.177).

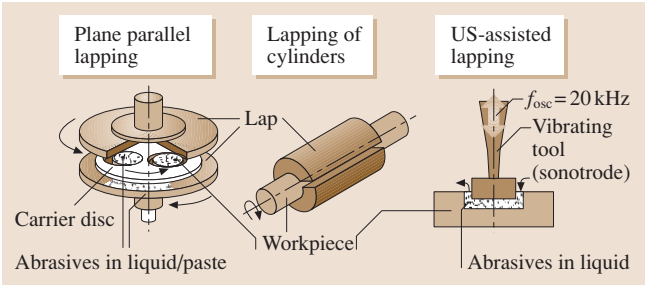


Fig. 7.184 Lapping processes

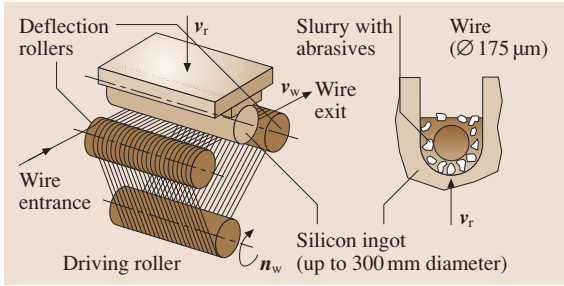


Fig. 7.185 Multiwire slicing

Due to the above-mentioned increase of silicon ingot diameters for wafer production a special variant of lapping has been developed called multiwire slicing by lapping (Fig. 7.185). In this process a silicon ingot is glued to a steel support and radially fed downwards to a multiwire field. Comparable to an egg-cutter, hundreds of wafers are separated at the same time, because on the specially developed high-strength wire a spray of silicon carbide abrasives in special slurry is applied. This process is capable of significantly increasing the productivity compared to ID cut-off grinding and currently works for diameters up to 300 mm. Parallel research is being done to develop an abrasive coated wire instead of loose lapping grains, thus reintroducing grinding for wafer manufacturing.

Abrasive Waterjet Machining. Abrasive waterjet (AWJ) machining is a material removal process with a high velocity slurry jet. The slurry is formed by the injection of abrasive particles into a waterjet by an orifice (Fig. 7.186) [7.124]. The abrasive waterjet head is controlled in two to five axes by a CNC system or by a robot arm. The process is mainly used for cutting a large variety of materials. The main advantages are the versatility regarding complex shapes and the ability to cut both ductile and brittle materials at very low process temperatures.

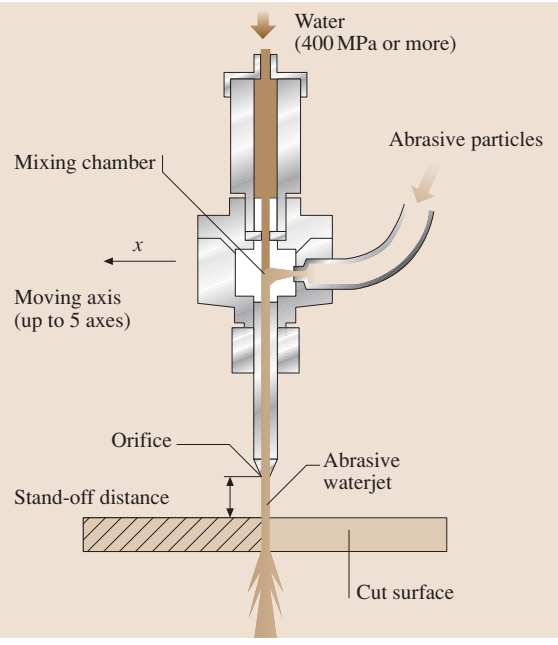


Fig. 7.186 Basic concept of abrasive waterjet cutting

Abrasive waterjet machining is usually performed at jet pressures up to 400 MPa. The applicability and the advantages of the cutting process beyond 400 MPa pressure is currently being investigated [7.125]. The success of this process at increased levels of pressure depends on the endurance of the system components subjected to high alternating stresses, the avoidance of solidification of water at ultra high pressures, and the good coherence of the jet.

The quality obtained in waterjet cut is higher than that achieved by conventional cutting processes, mainly in the edges. The dimensional and form quality will depend on the CNC system and the coherence of the jet. A cutting thickness of 200 mm and dimensional tolerances of about 0.1 mm can be achieved with this process. The main quality disturbances are related to the jet distortion or turbulence during cutting. Shape distortions occur when cutting corners, thick parts, or at high feed speeds.

The equipment for AWJ can be very simple and relatively inexpensive. A stationary waterjet is usually used in the aerospace industry to trim composites similar to a band saw. But abrasive waterjet cutting can also be performed in CNC machines and robots allowing the cutting of complex shapes. In any case, the waterjet has to be caught in a suitable catching system after cutting through the workpiece material to avoid danger to the operator or the environment. Furthermore noise protection measures are recommended.

7.3.3 Nonconventional Machining Processes

The nonconventional machining processes can be classified according to the type of energy employed: mechanical, thermoelectric, electrochemical, mechanical, and combinations. The way of material removal varies and so do the transfer media.

Mechanical energy is used in the abrasive-jet and waterjet applications, which have already been covered in Sect. 7.3.2.

Thermoelectric energy is used in electro-discharge machining (EDM) for making cavities mostly in difficult to machine materials. Metal removal is accomplished through the controlled vaporization of the workpiece material by suitably applied electric sparks. A very important variant of EDM applies thin wires to generate small gaps in the workpiece. This is called spark electro-discharge machining (SEDM), spark erosion machining (SEM), or wire-EDM.

The principle of electron beam machining (EBM) is the transformation of the kinetic energy of high speed electrons into thermal energy as they strike the workpiece. Laser beam machining is based on the transformation of light energy into thermal energy. Plasma arc machining (PAM) and ion beam machining (IBM) utilize ionized plasma for energy transfer. These beam machining processes are being used in hole-making and slicing operations and in welding metals and alloys.

Electrochemical energy is harnessed in electrochemical machining (ECM) and in combination with grinding in electrochemical grinding (ECG). Another application is the combination of ECM with electro-discharge machining (EDM), in the form of electrochemical-discharge machining (ECDM). The beam machining group also applies electrical energy in various forms.

Common in all nonconventional electrophysical-chemical machining is the application of high energy density in the range of 10^3–10^8 W/cm^2. High energy concentration is achieved by space and time localization of the emitted energy [7.126, 127].

Electro-Discharge Machining (EDM)

In the EDM process, material is removed by a series of discrete electrical discharges (sparks) in the machining gap between the electrode and the workpiece, as illustrated in Fig. 7.187. The dielectric fluid creates a path for the discharge as the fluid becomes ionized between the nearest opposing points. The initiation of discharge takes place to cause ionization and current starts to flow. As conduction continues, the size of the discharge column increases and so does the current. The discharge temperature is far beyond the melting point of the work materials, including exotic materials.

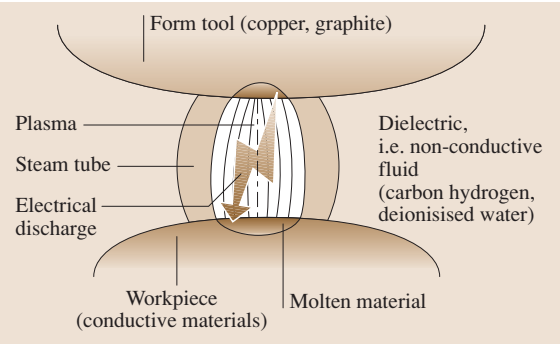


Fig. 7.187 Schematic diagram of electro-discharge machining (EDM)

However, the machinability of metals and alloys depends on their thermal-physical properties and the electric parameters of the process. Material removal is not affected by the mechanical properties of workpiece materials.

Electric-spark machining is characterized by use of electric discharge with an on-off time ratio, where the current amplitude to pulse duration is high. The tool-electrode is of direct polarity (cathode), applying powers between a few watts and several kilowatts. Application areas include cutting components of complex form by wire-electrode, broaching, and die sinking of precision components from refractory alloys, sintered carbides, and nonferrous metals with surface sizes up to 3–5 cm^2.

In electric-pulse machining the use of electric discharge with a small on-off time ratio with reduced current amplitude to pulses is common. The tool-electrode is of reverse polarity (anode). Application areas include the generation of complex surfaces with an area of up to hundreds thousand mm^2.

The dielectric fluid is an important variable in the EDM process. The methods of its application are shown in Fig. 7.188. It has three main functions: it works as an insulator between tool and workpiece, it acts as coolant, and as a flushing medium to remove chips [7.128]. It also affects electrode wear, material removal rate, and other EDM characteristics. The dielectric fluid normally used is hydrocarbon-based, but other fluids like tri-ethylene glycol and tetra-ethylene glycol are also used to improve the material removal rate. For wire erosion it is common to use deionized water in constant supply to the working zone.

Tool electrodes are made of graphite and copper, but other materials can also be used. The machined surface shows no direction of marks; on the contrary, the surface is formed from a number of craters and might show areas of molten material with accompanying surface integrity changes, as can be seen in Fig. 7.189 [7.129]. The application areas of EDM are shown in Fig. 7.190.

Electro-discharge machining provides machining accuracies of up to 0.005–0.02 mm in the case of through holes and up to 0.01–0.1 mm when machining cavities [7.130].

Surface roughness when machining through holes is in the region of $R_a = 0.4$–1 µm and $R_a = 1$–2.5 µm for machining cavities.

The material removal rate depends on the volume of material removed by individual sparks and by the discharge frequency. The volume of material is a function of the discharge energy, which increases with the cur-

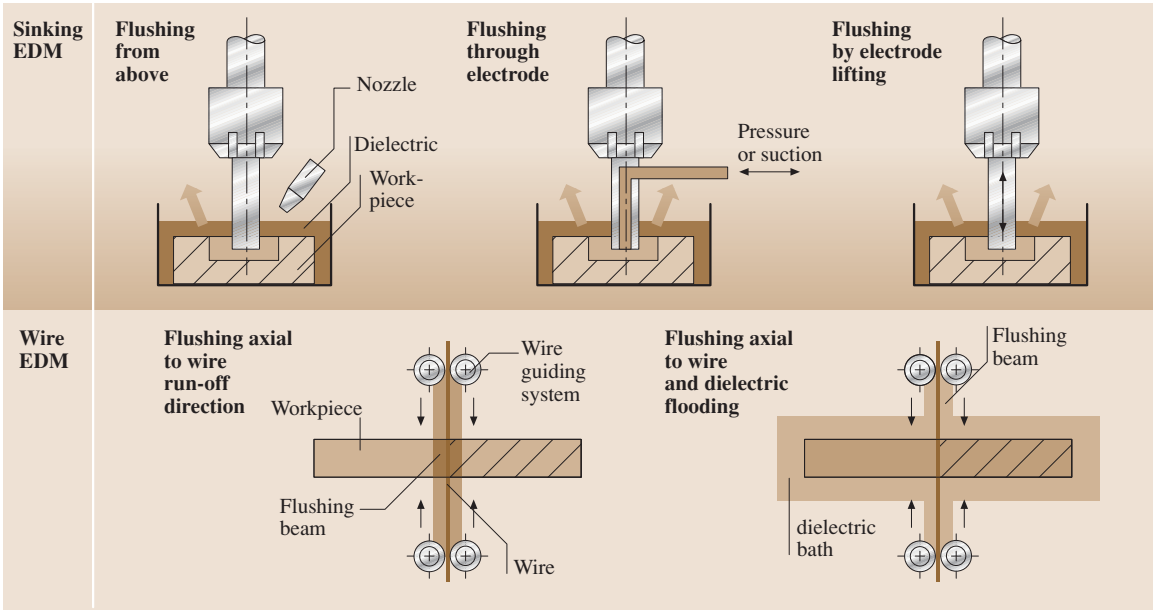


Fig. 7.188 Application of dielectric fluid in EDM (after [7.126])

rent. The typical removal rate when machining steels lies in the region of $10-12\times10^3$ mm^3/min.

Typical output parameters of wire electro-discharge machining (WEDM) are: dimensional accuracy 0.005–0.03 mm, surface roughness $R_a=0.4-2$ µm; the surface division rate for steels up to 300 mm^2/min, for carbides up to 120 mm^2/min.

The given parameters of electro-discharge machining allow these processes to be applied in tool manufacturing and in the production of components made from hard-to-machine materials with complex shapes.

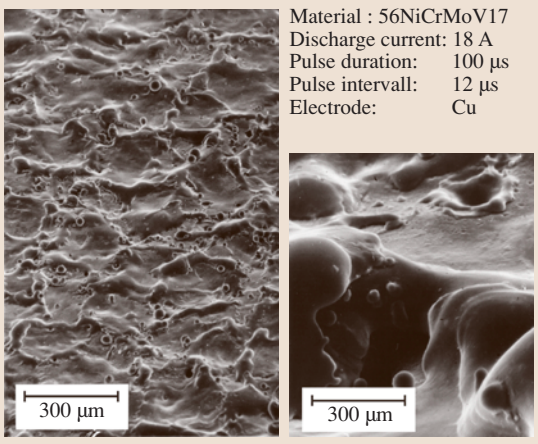


Fig. 7.189 Electro-discharge machined surface

Electrochemical Machining

The operating principles of electrochemical machining (ECM) have been known for a long time. Its two conductive poles are placed in an electrolyte bath and energized by direct current; metal may be removed from the positive pole (anode), the negative pole (cathode) is also metallic.

ECM requires a cathode tool with an appropriate mirror image of the cavity to be made; a workpiece and locating/holding fixture in close proximity to the tool, a suitable electrolyte supply and removal to/from the gap, DC electrical source with sufficient power to provide high current density between the tool and workpiece [7.131, 132]. The metal is removed rapidly from the workpiece, therefore provision has to be made to feed the tool to maintain the gap at a constant value.

Metal removal is governed by Faraday's and Ohm's law. The material removal is proportional to the current flow through the gap and the elapsed time, while the current is proportional to the applied voltage and inversely proportional to the gap's electrical resistance.

ECM is carried out as anodic dissolution of metal, and the product is removed from the working zone (gap) by an electrolyte flowing at speeds of about 5–50 m/s.

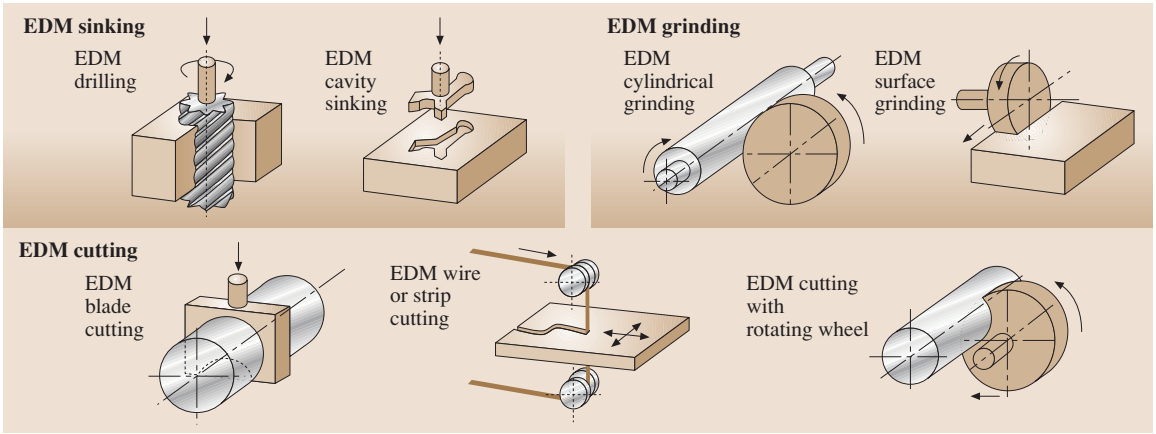


Fig. 7.190 Examples of electro-discharge machining

A DC generator supplies the current at a working voltage of 3–24 V governed by the material and technology requirements. A working gap between the two electrodes is maintained accurately by a process monitoring system between 0.02–0.5 mm. The tool-electrodes are typically made of stainless steel, brass and graphite. A typical ECM arrangement is schematically shown in Fig. 7.191.

Many process parameters affect the output parameters, including tolerances of the product. These can be electrical, chemical and mechanical parameters, which have to be monitored and controlled all the time.

The applications of ECM include turning, die-sinking, hole drilling/multiple-hole drilling, trepanning, broaching, milling, and deburring. In ECM sinking and broaching the shape of the tool-electrode is copied into the component. This is used to make cavities for dies, forging stamps, compression molds, casting forms, etc. The roughness of the machined surface is within the range of $R_a = 0.25$ up to $R_z = 20\,\mu m$. Feed rates of 0.03–0.5 mm/min can be selected.

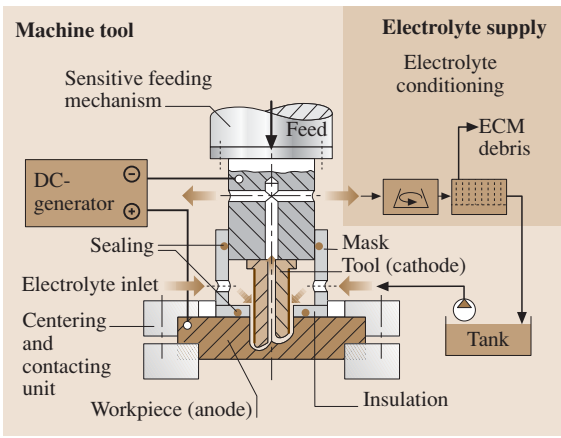


Fig. 7.191 Principle of electrochemical machining (ECM) (after [7.126])

Electrochemical milling, using rotary-cathodes, can be used for the production of profile, flat and round external surfaces by the cathode wheel at a removal rate of 150–200 mm^3/min in the case of stainless steels and 60–80 mm^3/min when machining carbides.

This process can be used to produce thread-cutting dies, various form tools, thread rolling dies with surface roughness of about $R_a = 0.63\,\mu m$ to $R_z = 20\,\mu m$. The process can also be used for machining magnetic materials (permanent magnets).

Electrochemical deburring is particularly suited for the deburring of holes, gears, and components of hydraulic equipment mainly in mass production.

Ultrasonic Machining

Ultrasonic machining (USM) utilizes a tool, which oscillates at high frequency beyond the audible limit (hence the name of the process), i. e. beyond 20 000 Hz. In the case of ultrasonic assisted lapping with loose abrasives, the tool has the same shape as the required cavity in the workpiece (see the section *Other Processes: Inside Diameter Cut-Off Grinding, Lapping, and Abrasive Waterjet Cutting*). The high speed alternating motion of the tool drives the abrasive grains across a small working gap against the workpiece. The impact energy is responsible for the material removal. Although this process is not restricted to brittle materials, and ductile failure works in machining tough materials as well, it is usually used to machine hard

and fragile materials, such as glass, ceramics, silicon, ferrite, ruby, sintered carbide, diamond and the like [7.133]. Besides the oscillation from the tool side, ultrasonic movement of the workpiece is also possible [7.134].

It is accepted that material removal results from the combined effects of hammering (impacting) abrasive particles in the work surface, the impact of free abrasive particles on the surface, cavitations, erosion, and the chemical action of the fluid employed.

The most important process input parameters controlling the material removal rate, surface roughness, and accuracy are the frequency and amplitude of oscillation, abrasive particle size, and by implication the impact force. During machining abrasive grains enter the machining zone as abrasive suspension.

Mechanical vibrations of the tool with ultrasonic frequency are achieved by a suitable electro-mechanical converter. Usually, the converter will consist of magneto-resistive elements with the ability to change their linear size in a variable magnetic field. In some cases, piezoelectric converters are employed in a variable electric field.

Tools for USM are made of structural steels, whilst carbides, CBN, silicon, and diamond are used as grits with grain sizes up to 3–10 μm [7.137].

The abrasive slurry moves into the machining zone either freely or under pressure and is removed by suction through the tool or workpiece, which substantially increases machining productivity. Mass concentration of grits in the abrasive slurry at free feed is in the region of 30–40% and 20–25% when it is supplied under pressure.

When machining sintered carbides, USM can be combined with electrochemical anodic dissolution. Processes of rough and finishing ultrasonic operations can be carried out on one machine tool.

USM can provide a removal speed of up to 5500 m^3/min when machining glass and of up to 500 m^3/min when machining carbides; the corresponding surface finish is $R_a = 0.32–0.16$ μm.

In recent ultrasonic developments the loose abrasive set-up is replaced by a tool with bonded abrasives (see the section *Grinding with Rotating Tools*). Regarding the kinematics it is an ultrasonic assisted grinding process, but in some cases it is referred to as ultrasonic milling. Besides milling and drilling also ultrasonic assisted turning has gained attention due to the improved wear behavior of the chosen tools [7.138, 139].

Beam Machining

Thermoelectric processes utilize concentrated thermal energy to remove material and electrical energy, in some ways, to generate thermal energy. The main characteristics of these processes are high temperatures and high thermal energy densities that can be achieved for material removal up to 10^9 W/cm^2. The main beam machining processes are: laser beam machining (LBM), electron beam machining (EBM), and plasma beam machining (PBM). Beam machining (BM) can be used for machining both electrically conductive and nonconductive materials.

Table 7.46 Characteristics and application of thermal sources (after [7.135, 136])

Thermal source	Limiting concentration of energy (W/cm^2)	Energy source	Application
Gas flame	8×10^2	Jet of the heated gas $T \approx 3500$ K	Cutting off, accompanying heating, maximum thickness up to 3 mm
Arc plasma	6×10^3	Gas and the metal steam ionized by electric discharge	Cutting off up to 3 mm, welding, heat treatment, welding
Electron beam	10^5	Electron beam in vacuum	Cutting off, welding (up to 20 mm/pass), heat treatment, welding
CW type laser beam	10^9	Beam of photons in a gas	Welding (up to 10 mm/pass), heat treatment, welding on, evaporation of layers
Pulsed laser beam	10^{10}	Beam of photons in a gas	Evaporation of layers, drilling of apertures, surfaces amorphous, shock hardening

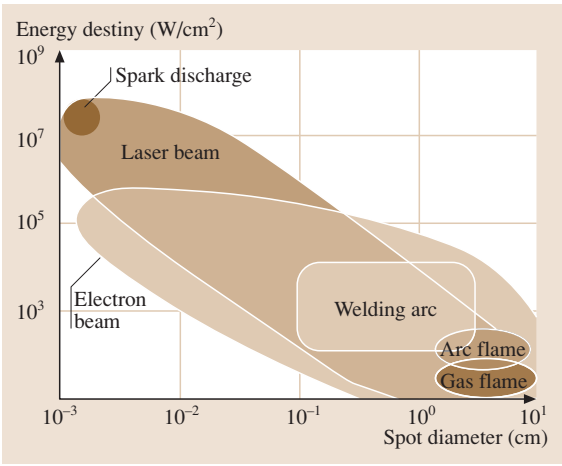


Fig. 7.192 Energy density of various thermal sources (after [7.135])

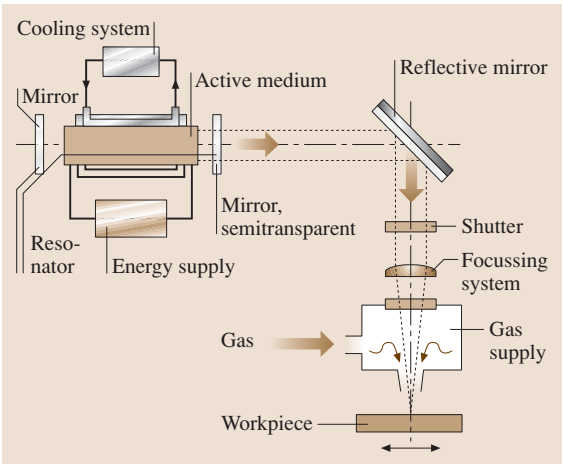


Fig. 7.193 Schematic diagram of laser beam cutting

The energy concentration of beam sources and that of comparative technologies are illustrated in Fig. 7.192. Of the beam-based technologies laser and electron beams provide the greatest intensity of energy in the region of 10^8–10^9 W/cm^2. The main characteristics of high density beam thermal sources are given in Table 7.46.

Laser Beam Machining. The technical term *laser* is the acronym of light amplification by stimulated emission of radiation. Laser cutting (LC) is based on use of the monochromatic electromagnetic radiation generated by the laser. The laser beam is formed by means of an optical system and is focussed on the machined surface, causing heating, fusion, evaporation, or explosive destruction of the workpiece material (Fig. 7.193).

In the radiation zone the form and diameter of a light spot changes over a wide range from several up to hundreds of micrometers. Depending on the temperature and energy concentration on the machined surface, various laser machining processes can be applied, e.g. drilling, cutting, marking, welding, heat treatment, and others. The application of laser beam (LB)-based manufacturing processes is summarized in Fig. 7.194. Meanwhile the application of laser technology is state-of-the-art in every area of manufacturing processes.

Laser cutting can be carried out in air, vacuum, or the required gas media. The type of LB-radiation is determined by duration, frequency of following and peak capacity, and also in a kind of continuous or quasi continuous radiation with the set average capacity modulated with a frequency of 5–50 kHz.

Laser cutting of metallic materials requires intensities of $> 10^6$ W/cm, which are achieved by focusing the laser beam with the aid of lenses or mirrors. The thermal material removal process, which is directed into the depths of the material, produces a kerf in the material when the feed motion is applied [7.140, 141].

The material is melted (laser fusion cutting), burnt (laser flame cutting), or vaporized (laser sublimation cutting) at the focal point of the laser beam, depending on the intensity and the length of interaction. It is expelled from the kerf by a stream of gas emitted from a nozzle coaxially to the optical axis (Fig. 7.193). The cutting gas also serves to protect the sensitive focusing optics from spattered material.

In laser flame cutting, oxygen or oxygen-rich gas is used as cutting gas, at higher cutting speeds. However, this leads to oxidation of the cut surfaces owing to the introduction of additional exothermic energy. In contrast with the other laser cutting processes mentioned above, inert gases (e.g. argon, nitrogen) are used as cutting gases, which results in a slower cutting speed. However, they produce an oxide-free cut.

The relative motion between the laser beam and the workpiece that is that produces a continuous kerf is achieved in various ways. For laser cutting of small, easily handled components, the latter are generally moved underneath the stationary laser beam. e.g. with the aid of an X–Y coordinate table. For laser machining of larger workpieces, the laser unit including the cutting tip is either moved across the stationary workpiece, or a movable system of mirrors is guided together with the cutting tip (*flying optics*) between the fixed laser unit

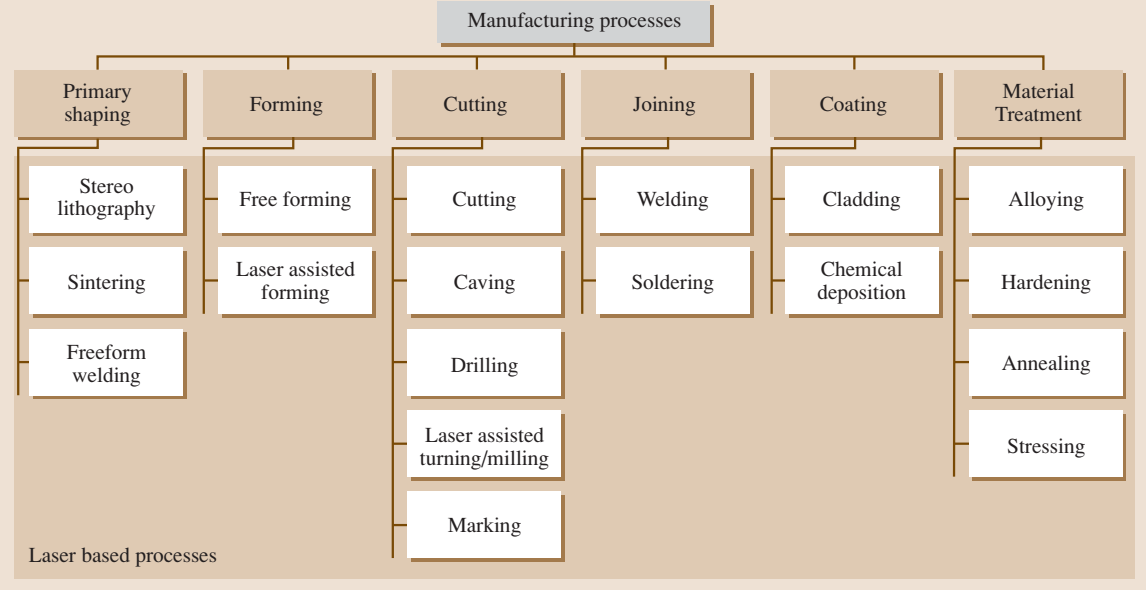


Fig. 7.194 The application of laser beam-based manufacturing processes

and the workpiece. For Nd:YAG lasers alone, flexible optical fibres may be used for guiding the beam.

The cutting speeds for other metallic and nonmetallic materials for a CO_2 laser with a power of 500 W are summarized in Table 7.47.

Table 7.47 Machining parameters for laser cutting of various materials (after [7.142])

Material	Thickness (mm)	Cutting gas/ pressure (MPa)	Cutting speed (m/min)
PMMA	4	Air/0.06	3.5
Rubber	3	N_2/0.3	1.8
Asbestos	4	Air/0.3	1.6
Plywood	3	N_2/0.15	5.5
Cement asbestos	4	Air/0.3	0.8
AlTi ceramic	8	N_2/0.5	0.07
Aluminum	1.5	O_2 /0.2	0.4
Titanium	3	Air/0.5	2
CrNi steel	2	O_2/0.45	1.9
Electrical sheet	0.35	O_2/0.6	7
Grey cast iron	3	N_2/1	0.9
Laser CO_2/500 W; focal length $f = 127$ mm			

High-powered lasers of the type mentioned above generally belong to laser (protection) class 4, which has the highest hazard level (except for laser machining systems with dosed working chambers, which are equipped with additional safety facilities such as interlock systems and radiation-absorbing protective windows). Allocation to this safety class means that even diffusely reflected laser radiation is a hazard to the skin and the human eye.

Technological laser systems mainly use the following types (active media) of laser radiation: excimer

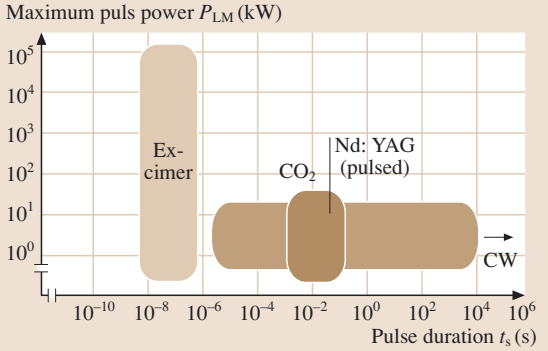


Fig. 7.195 Characteristics of different laser sources (source: LZH)

Table 7.48 Parameters and application area for various laser systems (after [7.136])

Laser type	CO_2	Nd:YAG	Excimer
Average power (W)	Up to 10^4	Up to 300 (pulse type) Up to1000 (continuous type)	< 0.3
Wave length (nm)	10 600	1064 532 (double frequency)	193, 248 308, 351
Pulses duration (s)	10^{-5}–10^{-2}	$(4–20)\times10^{-8}$	$(3–30)\times10^{-9}$
Pulses energy (mJ)	–	Up to 18	Up to 15 000
Pulses frequency (kHz)	Up to 20	Up to 50	Up to 1
Pumping source	Gas discharge	Lamps/diodes	Gas discharge
Machinable work-piece materials material	Glass, Al_2O_3-ceramics	Si_3N_4/AlN-ceramics, various metals	Polymers, plastics, ceramics, glass
Application	Cutting out, welding, heat treatment, drilling	Cutting out, welding, heat treatment, drilling, marking	Heat treatment, marking, surface modification

lasers (gas); CO_2 lasers (gas); Nd:YAG lasers (solid-state). The special characteristics of each laser system define its optimal field of application. The characteristics of the laser sources are shown in Fig. 7.195, and the main parameters are summarized in Table 7.48.

Basic elements of the equipment for laser machining are the generation source, beam guidance, forming and focusing system, and fixation and control of workpiece moving system (Fig. 7.196).

The laser equipment on the basis of solid-state lasers is generally used, basically, is used for precision machining in various materials (ceramics, polycrystalline glass, ferrite, ruby) and applications like e.g. apertures for input of a wire-electrode, apertures in thin foils and films, apertures in watches stone billets, diamond draw, atomizers, etc., or precision cutting out and marking.

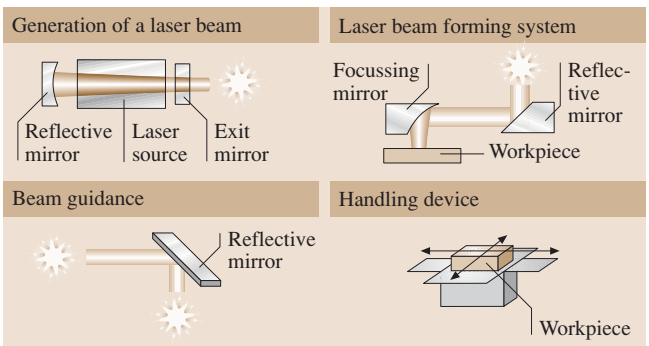


Fig. 7.196 Components of a laser manufacturing system

The fast development of laser sources and the large variety of beam forming systems open a wide field of applications as shown in Fig. 7.194. Especially in the fields of micromachining and medical applications laser systems have become an essential driver for innovation and improvement [7.82, 143].

Electron Beam Machining. Electron beam machining (EBM) is based on acceleration and focusing of electrons in a narrow bunch (beam), radiating from the cathode in deep vacuum with the help of a powerful electric field, to hit the machinable workpiece surface (anode). The physical essence of the EBM

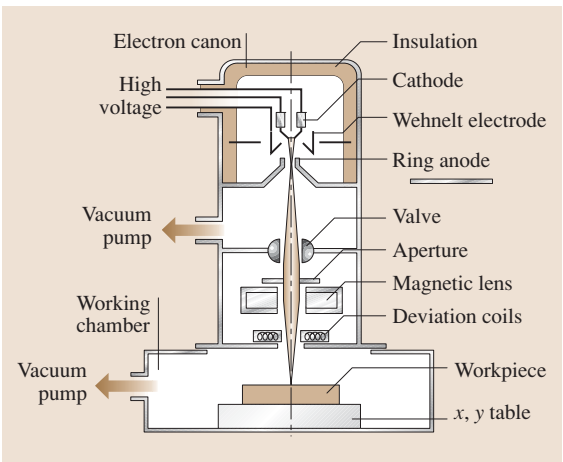


Fig. 7.197 Schematic set-up of electron beam machining

process is the transformation of electron kinetic energy into thermal energy. The electrons are emitted from a hot cathode and accelerated towards a ring shaped anode with a round opening (Fig. 7.197). The acceleration voltage is around 25–200 kV. The electrons reach the workpiece surface and release their high kinetic energy. The energy density of an electron beam is up to 10^5–10^7 W/cm^2, which is comparable to laser beams. Due to the fact that electrons, unlike photons, have a mass ($m_e = 9.108 \times 10^{-31}$ kg) and an electric charge ($e = 1.602 \times 10^{-19}$ A s), the energy transfer and effect on the workpiece surface is different compared to laser processes. To prevent electron collisions with gas molecules, the system is highly evacuated. The focusing of the electron beam is done with magnetic fields, which are generated by so-called magnetic lenses. The electrons can penetrate much more deeply into the workpiece material than laser beams. They transfer their energy via collisions with enveloping electrons.

EBM systems are more complex compared to regular laser systems. Due to the limitation of the working chamber this process is dedicated to special applications, where the benefits of high penetration depth, high impulse frequency, and fast beam deflection can be used [7.144]. Thus drilling is a perfect process to elaborate the potential of EBM, e.g. small bores (diameter down to few μm) can be made in thin foils at a rate of up to 10 000 bores/s without moving the target.

Typical application fields are combustion-chambers of aircraft turbines (difficult to machine cobalt-based material, several thousand holes) or spinning heads for glass fiber production (6000 holes of 0.8 mm diameter in 5 mm thick material [7.145]). A disadvantage of the technology is related to the fact that X-rays are generated above an acceleration voltage of 80 kV, which makes a severe shielding of the system necessary.

Plasma Beam Machining. Plasma beam machining is based on the use of low temperature open plasma, which is applied to increase operational properties of machined components such as wear resistance, corrosion stability, thermal stability, etc. Such an amelioration is carried out in order to attain the formation of functional coatings from corresponding materials, generated by a plasma jet, plasma welding, and plasma depositing. Furthermore plasma is used in some combined plasma-mechanical processes, in particular in plasma-mechanical machining.

Plasma coating is characterized by a great concentration of the thermal stream and high speed of the plasma jet. For coating fine grained powders are used (40–100 μm). The thickness of the deposited layer is around 0.3–0.5 mm and higher; deposition productivity is 2–4 kg/h.

Plasma cutting is characterized by local removal of metal along a cut line by a plasma jet using quality plasma forming gases like argon, nitrogen, hydrogen, air, etc.. It is applied for cut off stainless steels with thicknesses up to 60–80 mm and low-carbon and low-alloying steels with thicknesses up to 30–500 mm. After plasma cutting the surface roughness may reach $R_z = 80$–160 μm.

Combined Methods of Machining

For further development of manufacturing technologies a combination of energy sources is possible. With a complex joint use of mechanical, thermal, chemical, or electrical energy an enhanced material removal, better surface quality or improved tool life time can be achieved. Examples for these process combinations are e.g. electrochemical grinding, electro-discharge grinding, ultrasonic-electrochemical, electro-discharge-chemical, anodic-mechanical, plasma-mechanical, and laser-mechanical processes.

Electrochemical Grinding. Electrochemical grinding (ECG) is carried out by overlapping material removal by microabrasive grains (diamond, CBN) with anodic (electrochemical) dissolution. Anodic dissolution of metallic workpiece materials reduces the microchip thickness and the area of mechanical contact between workpiece and the grinding wheel [7.146]. Furthermore ECG reduces the material resistance against mechanical penetration by means of reduction of the strength of the superficial microlayer.

ECG processes work at a voltage of up to $U_p = 5$–10 V (at machining with independent electrode $U_p = 24$ V) and a current density of up to 15–150 A/cm^2. Nitrate/nitrite solutions are often used as working media (dielectric fluid). They contain various passivation additives (soda, glycerin, triethylamine, etc.) for reduction of the corrosion activity.

ECG processing is applied on surface ground components of hard, magnetic, heat resisting steels and alloys; surface and cylindrical grinding of thin-walled, nonrigid components; profile grinding; grinding of viscous materials, etc.

Electro-Discharge Grinding. During electro-discharge grinding (EDG) metal removal is carried out by microcutting of bonded abrasive grains while the grinding wheel working surface is continuously influenced by electroerosion. Electric discharges provide an opening of the grinding wheel topography, allowing new abrasive grains to come into contact, and the removal of adhering chips from the wheel surface to prevent loading. Also just cut chip segments might be directly evaporated in the zone of contact. Discharges occur between the workpiece and the wheel, or between the tool and a specially adapted additional electrode [7.147, 148]. EDG processes can be actively controlled and their intensity can be adjusted to provide substantial increase and stabilization in the lifetime and cutting ability of the grinding wheel.

For EDG processes either standard cutting fluid or 3% water soda solution is used as working medium. Current-conducting metal bonded abrasive wheels (or wheels with diamond or CBN grains) are connected to the positive pole and the workpiece is connected to the negative pole of a pulse voltage source. EDG is applied in tool grinding, surface grinding, and cylindrical internal and external grinding machine tools.

Laser-Assisted Machining. Laser-assisted machining (LAM) has found rising acceptance due to its potential to significantly increase machining efficiency; in particular, in processes of punching and cutting complex shaped workpieces. Laser radiation is employed for two reasons. Firstly, the laser source is used for heating (thus annealing or hardness reduction) of the workpiece material surface layer directly in front of the cutting tool. By this tool lifetime is extended and the productivity is drastically increased. It even opens the possibility of applying cutting technologies to materials that were not machinable in this way before, like laser-assisted turning of ceramics [7.149]. Secondly the laser can be used for final surface formation (for example, a groove after milling by an end mill). Applied in this combination LAM processes increase the accuracy and quality of the machined parts.

7.4 Assembly, Disassembly, Joining Techniques

Considering manufacture of tomorrow, production process and technology of joining will continue to hold its dominating position worldwide in the production of added value. Figures 7.198 and 7.199 clearly show the increasing economic importance that must be attributed to joining in those industrial branches where it already has been intensively utilized so far – an importance as is hardly attributed to any other metal working manufacturing procedure. A large number of bonding methods having evolved within the last one hundred years have gained in importance by developing from a conventional into a high-performance welding procedure. Highly effective joining technologies have been adapted to the specific characteristics of material and structural parts and developed into microjoining or hybrid joining techniques.

Fig. 7.198 Value added by joining industry (2003) [7.150], with permission

Today, joining technology in general covers about eighty different modern joining techniques which are in use worldwide, even though with varying intensity.

Therefore it cannot be aim of these chapters to provide the international readership, engineers, specialists or students with process engineering descriptions of all joining procedures known. The author and his coauthors agree with the fundamental idea of this book to present even a profound description of joining technology, though with intended restriction to selected state-of-the-art technologies (Fig. 7.200).

Thus it seems to be quite reasonable, within the frame of this technology handbook, to point out trends in the chapters presented, and to show the interested reader how an innovative product optimized in its

properties can be produced dependent on innovative materials, modern principles of design and manufacturing techniques (Fig. 7.201).

Pointing out trends in joining technology offers, on the one hand, the possibility to introduce current state-of-the-art technologies and to describe them in a competent way; on the other hand, the reader may draw conclusions from these high-technologies about the conventional joining processes. With that, authoritative standard works and bibliographical references listed in the appendices to each chapter will provide assistance.

Considering the production processes of tomorrow, the entire range of joining-technological processes as described here will certainly develop into centres of initiation and growth for a rapid and smooth manufacture, and will play an outstanding role in all phases of the so-called production life cycle, i. e. from design and development of a product through all phases of its manufacture and maintenance to recycling (Fig. 7.202).

Joining technology today is not only based on a substantial range of conventional process technologies but has also derived high-technological procedures from these and is, besides, ever more developing towards microjoining technology or even a nanotechnology.

History in the past and present reveals that joining technology is and has always been continuously expanding into new fields of practice, as a view over 2000 years back to Roman times (soldering) shows. Thus, at present, modern joining technology is increasingly engaged in supporting further development of modern organ- and vessel surgery.

7.4.1 Trends in Joining – Value Added by Welding

During the development of new technical products and the further development of existing technical products, greater attention is being paid to optimized material utilisation which also takes account of local stresses. Even individual components are being composed from different materials to an increasing extent. Examples are tailored blanks made of different steel sheets with various thicknesses and surface qualities or components which are coated locally against corrosion or wear. Furthermore, combinations of steel with light metals or plastics are being used ever more frequently in lightweight construction and the significance of nickel-base alloys is continuing to grow in the construction of power stations and aircraft engines.

Fig. 7.199 Value added by joining industry in different sectors (2003) [7.150], with permission

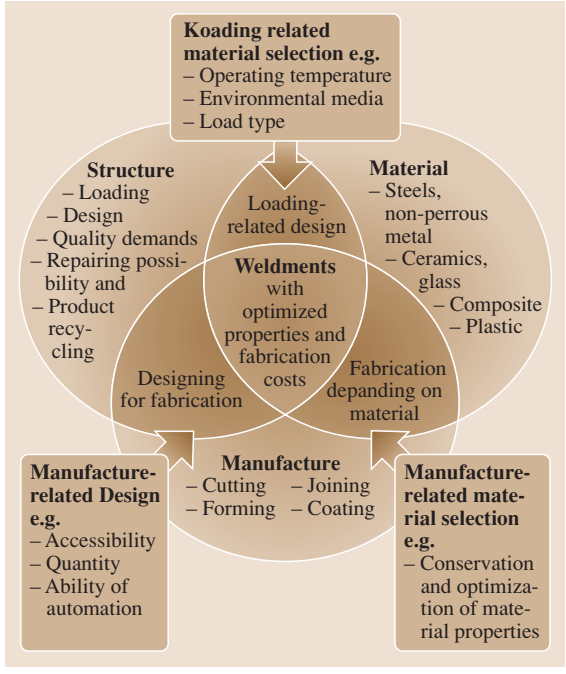


Fig. 7.200 Joining technological correlation between structure, material and fabrication

To an increasing extent, joining-technology processes will be necessary for this purpose and, in the future, will be core elements for quick, problem-free and reliable production – in this respect, they will be integrated into the chain of the manufacturing processes. Thus, joining technology remains economically significant; it is making a fundamental contribution to

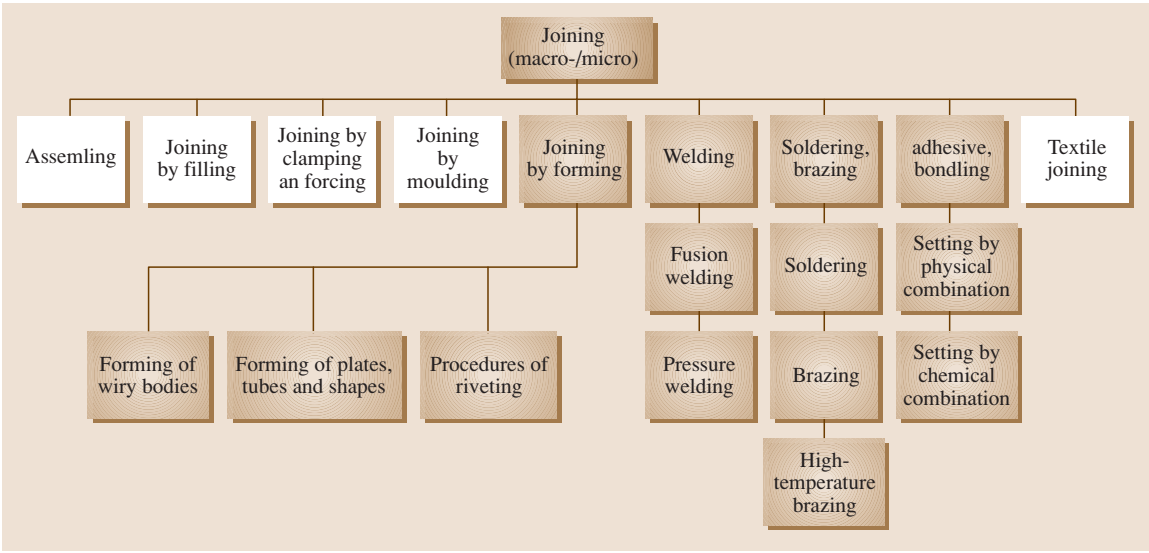


Fig. 7.201 Trends of joining procedures (DIN 8593)

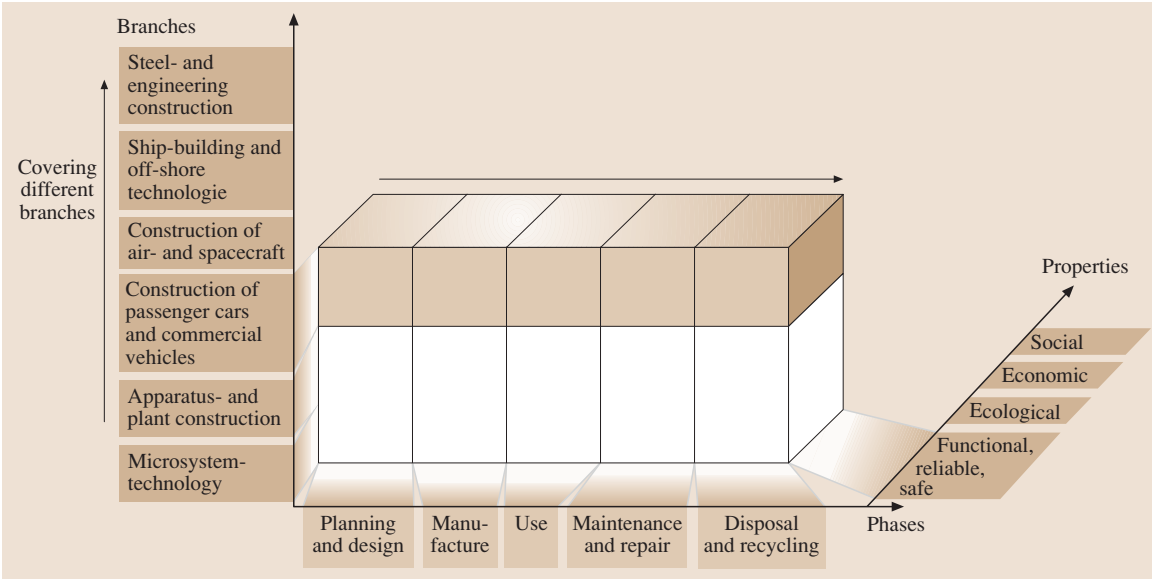


Fig. 7.202 Joining in product life cycle (source: DVS)

the value added by the producing sector and has good prospects of growth. Joining technology will play an exceptionally important role in all the phases of a product life – these phases encompass the design, development, fabrication, utilisation, maintenance and repair of a product as well as its recycling and disposal. Not only mechanisation and automation but also the use of industrial robots for joining will increase [7.151]. In this case, consistently interdisciplinary approaches in research, development, fabrication and application will be adopted in order to incorporate requirements resulting from the product, production and material development.

For economic reasons, the simulation of processes and properties is becoming the key factor also for the application and further development of joining pro-

cesses and materials. Those processes and materials which evade reliable simulation will no longer be used in the future. Simulated joining processes permit the virtual testing and optimisation of the joining processes even in advance of the production, permit the planning and implementation of quality-assurance concepts before the joining process and help during commissioning, operation and repair. At an early stage, they also provide information for the necessary training of personnel for the fabrication, maintenance and repair. The objective must be to permit a universal concept for the simulation of the entire process chain on the basis of the elements consisting of process simulation, structural simulation and material simulation [7.152].

The reliable calculation of the resistance, durability and strength of joined products when they are subjected to all the stresses to be expected is still of elementary significance for their utilization. Particularly with regard to cyclic stresses and the fatigue strength of joint structures, stringent requirements continue to be placed on design and calculation (Fig. 7.203).

This also results from the fact that joining technologies and joint geometries are being refined continuously. Another factor is that, in certain fields (e.g., in the transport sector), materials with a low volume weight and a higher strength are being used increasingly in order to meet the demands of lightweight construction. The further development of computing technology permits strength analyses also on the basis of local stresses. The aim is to reliably record fabrication influences on the strength of joined products. This encompasses an assessment of the effects of the joining technologies used and of the after-treatment processes used as well as the recording of the weld quality. It is necessary to determine the differences between the characteristics established on small specimens and the properties of products joined in reality. To this end, the existing database for fatigue strength must be analysed while paying attention to various aspects. It is indispensable to record residual stresses in a differentiated way and to take multiaxiality into consideration. Operational-strength analyses and fracture-mechanical assessments of joints are the state of the art. In many cases, requirements are defined by sets of rules and guidelines and must be developed further. The vision being pursued is that a combination of simulation, calculation and experiment results in material-dependent characterization which takes account of the real structure of the joint and of the actual properties. The objective is to exactly predict the behaviour of joined products when they are subjected to stresses. With regard to the strength, joints will no longer be regarded as weak spots.

Basically, it may be assumed that the definition of weldability until now (as a component property with the concepts of the welding possibility of the fabrication, the suitability of the material and the welding reliability of the design) must be refined by a definition of the joinability – with the subdivisions of the joining possibility, the joining suitability and the joining reliability [7.153] (Fig. 7.204). This terminology takes into consideration the fact that, in addition to welding-technology processes, other joining-technology processes such as brazing, adhesive bonding, mechanical joining and also coating are being used to an increasing degree.

In a continuous process, developments from information and communication technology will be introduced further into joining technology. Interface management will permit the communication between various information systems at various levels (Fig. 7.205).

Greater importance has been attached to repair and maintenance concepts especially for joined products in road-vehicle construction. With almost 45 million vehicles in Germany and around five million accident repairs and 30% seriously damaged body structures every year, economically viable repair concepts are

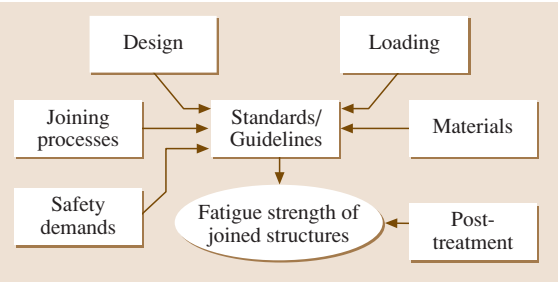


Fig. 7.203 Construction of joined metallic structures

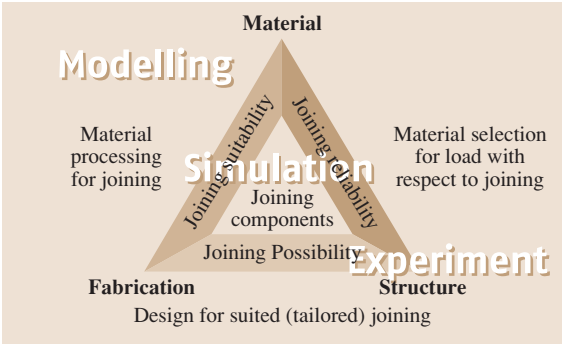


Fig. 7.204 Correlation between structure, material and fabrication

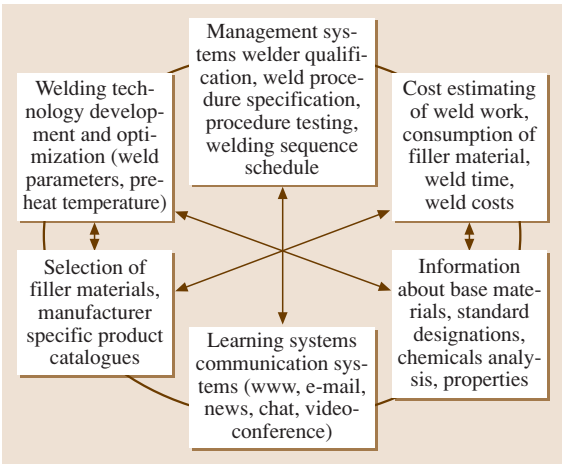


Fig. 7.205 Information- and communication systems – collaboration – interfaces

becoming more significant. Repair concepts must be an integrated part of the product development and must take account of the increasing utilization of high-strength steels, light-metal alloys, plastics and tailored blanks as well as of the applications of the laser-welding processes, adhesive bonding and mechanical joining. This is important because the joining processes applied in the production may have a direct influence on the possible repair costs of vehicles. Moreover, the costs of third-party and own-vehicle insurance are dependent not only on the accident frequencies but also on the repair costs. Joining processes which lead to rapid fabrication but generate high costs in the cases of repair may therefore be counterproductive. Modular construction methods which are suitable for repair are being applied especially in the case of vehicle construction. Concepts for qualified training in new repair techniques have been developed [7.154].

The people's justified wish for the protection of their environment is leading to particular expectations with regard to technology – this applies to joining technology as well. As examples, the requirements resulting from this can be described with the following headwords: Economical handling of primary and secondary energy, conservation of raw materials, avoidance or elimination or storage of residual materials, selection and utilization of reusable materials as well as recycling processes for joined components. In addition, measures for health and safety in joining technology remain urgent [7.155].

With a comprehensive concept incorporating the whole product life cycle, joining technology must react to these requirements specified here as examples. Such a concept is described in the following text which is preceded by a few trends from research and application in joining technology. For the future, a competition between the joining processes is emerging in the applications. Personnel qualification and competence management are becoming determining factors for the competitiveness of the companies in joining technology. As an example, this is shown on the basis of statements about the value added by welding technology in various welding-intensive sectors in Germany. The fundamental statements are summarized at the end and are connected with an outlook.

Trends in the Joining Processes

The statements in this section are essentially based on the results of the cooperative technical-scientific work in the DVS, findings from enterprises and research institutes are included.

One main focal point of the joining-technology research is to be found in the further development and optimization of joining processes, including automation. Consideration is given to the processes of welding, soldering and brazing, adhesive bonding and microjoining. In this case, welding accounts for the majority and is followed by the research in the areas of adhesive bonding, microjoining and soldering and brazing. It is significant here that the joining process of adhesive bonding has already become established as the second research field in joining. But also the future importance of the mechanical joining processes should not be underestimated. They are dealt with briefly below. Initiated by the trend towards lightweight construction and by other technological trends (e.g., the joining of coated sheets and the increasing modularization of the fabrication), there is resulting competition between the joining processes with regard to applications. Here, the thermal joining processes are competing not only with each other but also with low-heat and *cold* joining processes.

A more detailed analysis relating to the welding processes shows, that gas-shielded arc welding as the stock process in many companies continues to be an essentially determining factor in the research work. However, the beam-welding processes (laser beam welding and electron beam welding) already make up a substantial part of the present research work – with an upward tendency. In contrast, research work on resistance welding shows a slight downward tendency.

The necessity of using hybrid welding processes to a greater degree in practice is also shown in the research. This is reflected in the expectation of using these to a greater degree in practice. In this respect,

it is primarily the combination of laser welding with gas-shielded metal-arc welding which is regarded as a hybrid process. However, combinations of laser welding with gas-shielded tungsten-arc welding and with plasma welding are also being used. Basically, those welding processes in which different welding processes are coupled via a common molten pool and a common process zone thus exists are designated as hybrid welding processes. For the sake of completeness, attention is drawn to combinations of the following processes: welding and adhesive bonding as well as mechanical joining and adhesive bonding. These hybrid joining processes will possibly be important for the development in joining technology.

A relatively high proportion of research in the field of microjoining technology corresponds to the trend towards the miniaturization of products. Research trends with regard to the microjoining technologies show applications for example in high-temperature microelectronics and in optoelectronics. But also research focused on quality assessment dealing with simulation, calculation and reliability is a must. Microjoining processes under investigation are leser beam welding, soldering in electronics, bonding and adhesive bonding.

Not only the processes but also the materials influence heavily the research work in joining. Steel makes up to a significant amount of research, but light-metal alloys (aluminum and magnesium) will become ever more significant in the research and application. The already indicated rapid developments in lightweight construction are documented in the research contents. These are also associated with the increasing importance of *multimaterial design*, i. e. with the selective combination of materials with different properties, dimensions and surface conditions. As summarized on Fig. 7.206, highest-strength steels, light metals and mixed joints will be main focal points of the research in the future. This statement is confirmed not only by an extensive industry poll but also by applications in the companies.

Increasing significance is being attached to research projects in the fields of calculation, design and simulation – not only in the field of microjoining as mentioned above, but also in structural joining. There will also be more research in the area of health and safety. Points of research here are to be made not only to processes but also to the recording and assessment of welding fumes and to early-diagnostics methods in occupational medicine.

In this respect, it has also been proven that even greater consideration than until now should be given to research projects with an evident reference to the quality of the joined products (Fig. 7.207). Here, not only nondestructive testing but also sensor equipment and process diagnosis are emerging as main focal points – with the aim of controlling joining processes according to quality characteristics to an even greater extent than until now.

Coming back to the joining processes, it is possible, in simplified terms, to make the following trend statements about individual joining processes.

Gas-shielded arc welding retains a dominant position and the latest research supports the applications and further developments of these processes. However, an estimation for the future suggests that gas-shielded arc welding will not grow as quickly as in the past years.

The processes of resistance welding will hold a dominant position due to their process performance and productivity – anyhow they will lose applications, in particular because of the increased application of mechanical joining technology but also because of adhesive bonding. It remains to be seen, whether developments like hybrid processes (e.g., a combination

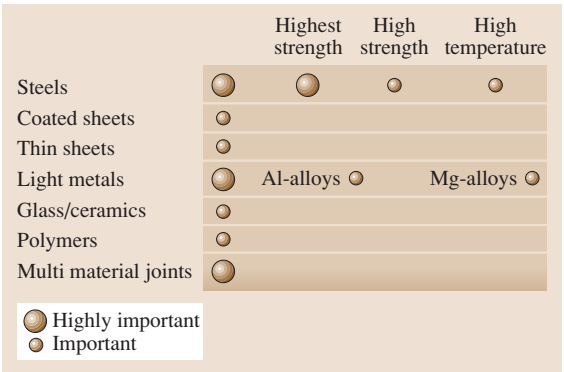


Fig. 7.206 Future research activities – materials

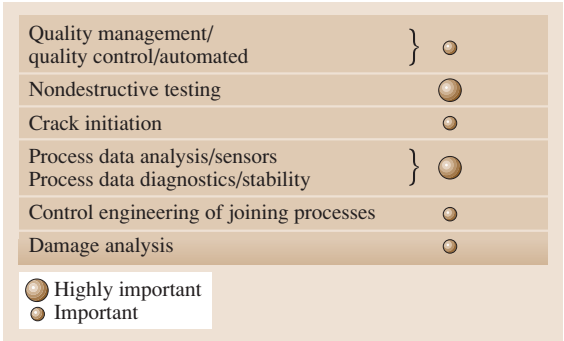


Fig. 7.207 Research in quality

of resistance spot welding and adhesive bonding) will prove its success.

The beam-welding processes (especially the laser beam welding processes) will become far more significant as a result of the rapid transfer of the findings from the research to the application. In this case, the growth of electron-beam welding will possibly slow down.

These statements are supported by the already specified study into the research needs of installation manufacturers and system suppliers [7.155]: the welding process with the highest growth in the next ten years from the viewpoint of the users will be laser beam welding – however, its growth will be accompanied by adhesive bonding, mechanical joining and soldering and brazing (Fig. 7.208).

For individual welding processes, a few examples of development potentials and research approaches are to be summarized below:

In the case of gas-shielded arc welding, the electronization of the power-source technology as well as the integration of fundamental operating functions into the welding torch will make further advances. Moreover, new process variants are being developed. Low-heat joining processes will be of particular importance for the future. Approaches are also being shown towards the development of an optimum installation concept for the user not only with easy-to-operate machines but also on multifunctional machines.

With regard to resistance welding, the servo welding guns are being refined and the operator interfaces on the welding controllers are being optimized. The goal is to perfect the sensor technology for on-line quality testing in the case of resistance welding as well.

For electron beam welding, improvements may be expected in the process observation and in the process recording as well as developments for the beam deflections. Appliance developments are optimizing vacuum electron beam welding by shortening the off-times. Non-vacuum electron beam welding is also the state of the art today.

For laser welding, the utilization of new beam sources is being tested and an improvement in the on-line process control is being striven for in the case of this process as well. The objective is to simplify the handling jigs by using flexible beam manipulators. Methods of beam deflection are leading to new applications in the form of the remote welding of large-surface, three-dimensional components. The utilization of welding filler materials is also being tested. Hand-guided laser systems are in use.

The monitoring and on-line regulation of the welding process as well as the on-line documentation of the weld quality are seen as technical challenges for all the welding processes. In all cases, top priority is being attached to the operator-friendly implementation of technical solutions in the appliances and systems.

Because of the diversity of the application possibilities in each individual case, these trend statements must remain in these general terms. In addition to the stable further developments in the specified processes, new welding processes are also being applied. Friction

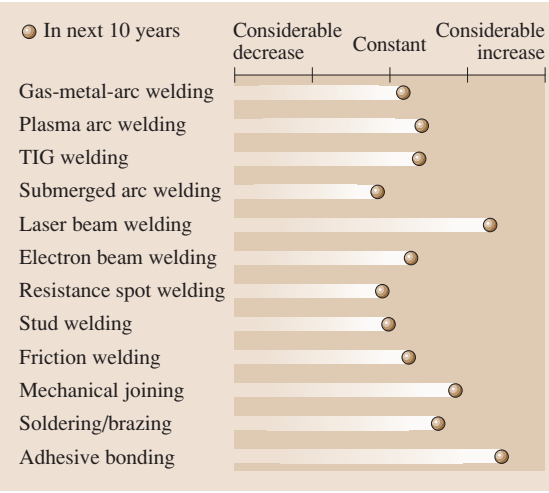


Fig. 7.208 Development of joining procedures from the companies point of view

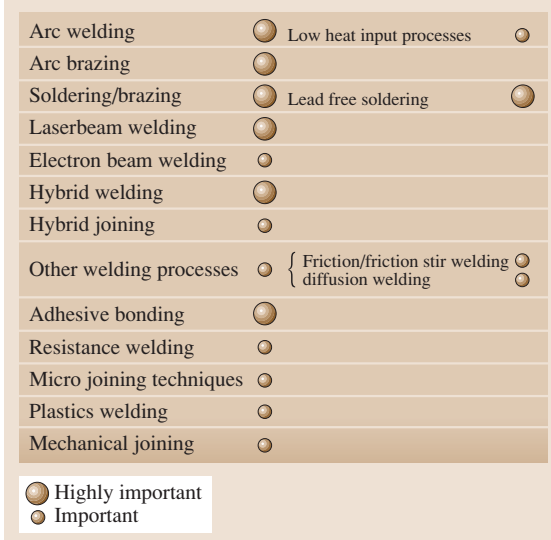


Fig. 7.209 Research/joining processes

stir welding in which it has been possible to exploit interesting economic applications after a relatively short development time is regarded as an example of this. Further examples are friction spot joining and magnetic-pulse welding. Figure 7.209 makes a rough assessment of the research fields of joining processes.

Before the explanation of the effects of the statements made here about the joining processes in some welding-intensive sectors, the value added generated by welding technology in Germany should be considered.

Value Added by Welding Technology in Germany

Studies conducted in 2001 [7.150] and in 2005 [7.156] show the economic importance of joining and welding.

- The production value of the companies providing goods and services for joining in Germany amounts to approximately Euro 3.6 billion. These goods and services include welding machines, appliances required for welding, welding filler materials, adjuvants, gases, protective clothing, testing machines and also the further training of the welding-technology personnel.
- The companies working for this production value employ 37 000 people (converted to full-time staff) and generate a value added amounting to around Euro 1.6 billion.

Investigations conducted by the DVS have indicated that the German market for ancillary supplies for joining technology may be estimated to be about one third of the European market and this in turn one third of the global market. For 2003, it is thus possible to estimate a value of approximately Euro 15 billion per year for the value added generated by ancillary supplies for welding technology all over the world.

However, the specified figures do not serve to define the value added by the companies which use joining technology in their fabrication in order to manufacture their products. It is far more difficult to estimate the economic benefit for this sector because welding is just a part of the value added in these companies and only few statistical values are available. A calculation is carried out here on the basis of a macroeconomic model. In this case, the study comes to the following results (Fig. 7.198): The macroeconomic value added by the production and application of welding technology in the investment-goods industry in Germany amounts to around Euro 27 billion (i. e. 4.8% of the value added by the producing sector in Germany). In this respect, around 640 000 employees are directly or indirectly connected with welding technology.

Results show that joining technology can hold its own even in a difficult economic environment and is proving to be a cross-sectional technology which, because of the close cooperation with various sectors, is relatively resistant to economic cycles and can open up new sales markets. A comparison of data from 2005 with the data from 2001 shows an increase in the value added of 18% with a simultaneous increase in the numbers of employees of 5% in welding technology. This is indicative of higher productivity.

This results from the strategies pursued by the companies in joining technology and by the companies applying it with the three following main focal points:

- Concentration on core competences
- Adaptation of the production processes
- Increase in the labour productivity

In Germany, a number of sectors are regarded as welding-intensive, including vehicle construction (road vehicles, rail vehicles, ships and aircraft) as well as metal construction and mechanical engineering. On average 5% of the total value added by these sectors is generated by welding technology (Fig. 7.199). A few joining-technology trends for these sectors are considered below.

Trends in Welding-Intensive Sectors in Germany

The following statements are essentially based on the results of the cooperative technical-scientific work in the DVS and make no claim to being complete.

Joining Technology in Road-Vehicle Construction. In order to reduce weight, tailored blanks are being used to an increasing extent in bodymaking. The combination of sheets with different materials, material thicknesses and/or surface conditions (as a rule, welded together using the laser beam) serves to achieve better material utilization with regard to the stress-bearing capacity. The tailored blanks allow the increasing use of high-strength ductile steels. The bodies are assembled not only by means of resistance spot welding and adhesive bonding but also by means of mechanical joining and gas-shielded arc welding. In this respect, the latter process is utilized in regions subjected to high and dynamic stresses. Another application is for the attachment of studs in order to fasten built-in and at-

tached parts according to the various methods of stud welding.

The joining processes are selected according to the criteria which are customary in other sectors as well. These are the joinability (regarded as a component property) and the economic viability. In this respect, it must be borne in mind, precisely in large-scale series fabrication, that the cycle times of the fabrication lines are very short. This necessitates the utilization of high-productivity welding processes.

In the last few years, the efforts to reduce the weight of vehicles of all kinds with the aim of reducing the fuel consumption during operation, the development of new materials and the developments of the joining processes have led to fundamentally new strategies for the construction of road vehicles. In addition to steel, light metals and plastics are being used to an increasing extent wherever this contributes to weight reductions without affecting the properties of the end product. This has resulted in pure aluminum bodies or in bodies made of a material mix. For such bodies with a structure consisting of different materials (multimaterial design), the selection of the correct joining technology constitutes a particular challenge.

Not only mechanical joining (bolting, clinching and riveting) but also laser welding and MIG welding are used for aluminum bodies. Resistance spot welding is now applied to a limited extent only. Thermal but low-heat joining processes which, for example, get around the difficulties relating to the metallurgical compatibility when different materials are welded are being developed for these applications as well. It remains to be seen how this *competition* between the thermal, low-heat and *cold* joining processes develops (Fig. 7.210).

Adhesive-bonding technology is becoming ever more significant for the assembly of bodies not only made of steel but also made of aluminum or material combinations because linear structural adhesive-bonded seams produce better stiffness behaviour than the punctiform resistance weld spots and seal the joint at the same time (Fig. 7.211). The disadvantage of this solution is that, as in the case of various welding processes, the sheet must overlap in the joining region and adhesive-bonded seams only obtain their strength at the end of the fabrication process during the firing of the paintwork. Other additional joints (manufactured, for example, by means of spot welding or clinching) must then ensure the necessary cohesion of the body during the fabrication.

The search for economically viable solutions is continuing: preference is given to the use of butt welds in bodymaking as well. Today, this is feasible and economically possible with laser welding since the fabrication of the raw parts is constantly improving its compliance with those tolerances of the single parts which are necessary for this purpose. The achievements with regard to the further development of high-power diode lasers for welding precisely in the thin-sheet range are giving cause for great hopes on the cost side as well. Where this is not successful, hybrid processes (e.g., the combination of gas-shielded arc welding and laser welding or laser brazing or gas-shielded brazing) are being used as alternatives.

Joining Technology in Rail-Vehicle Construction. As in motor-vehicle construction, lightweight construction is a main focal point in rail-vehicle construction as

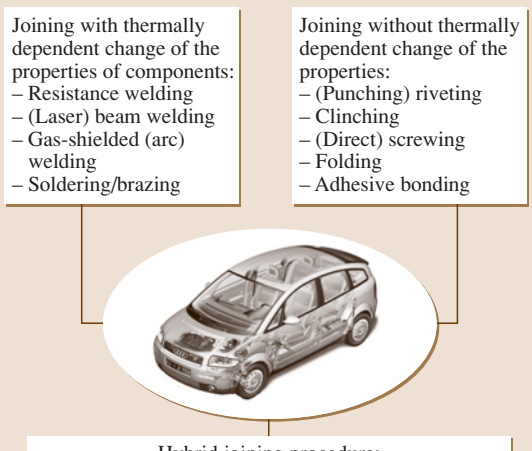


Fig. 7.210 Vehicle construction as *welding sector*

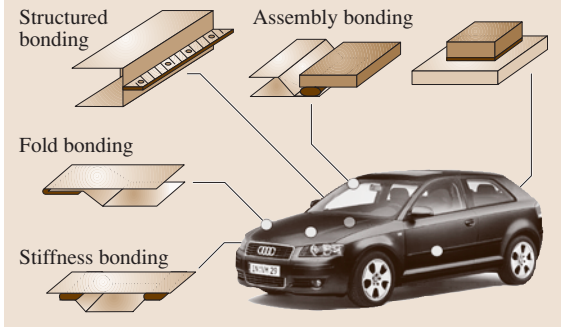


Fig. 7.211 Bonding systems

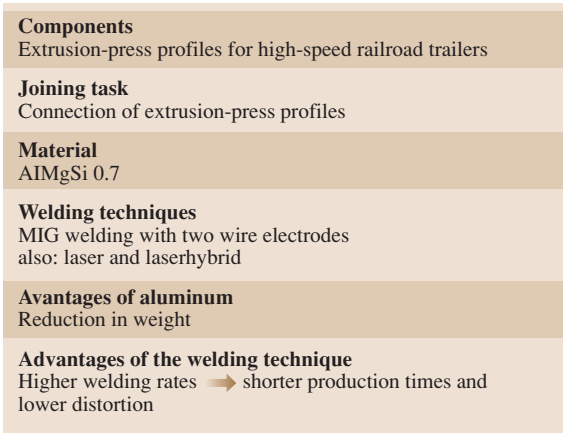


Fig. 7.212 Joining of aluminum materials through welding

well. The utilization of aluminum and its alloys for the construction of passenger carriages for the trains is gaining increasing acceptance. The joining technologies are, in part, the same as in motor-vehicle construction: gas-shielded metal-arc welding (also as twin-wire gas-shielded metal-arc welding, as shown on Fig. 7.212), laser welding and mechanical joining. Furthermore, rail-vehicle construction is developing into an application field for adhesive-bonding technology. However, since the components to be joined are thicker than in the case of the road vehicles, friction stir welding can be used as a young joining process, in particular for the very long longitudinal welds when extruded sections are welded together, for example, for the manufacture of the carriage roofs. However, the trend towards the utilization of aluminum is certainly reversible. Also high-strength steels and possibly even once again high-alloy steels will be used but the bogies will continue to be manufactured from steel with the aid of arc welding.

Joining Technology in Metal Construction. In both building construction and civil engineering, metal structures (as a rule, made of steel) are welded using arc processes. The degree of mechanization during welding is stipulated depending on the repetition frequency of the individual parts or of the weld lengths. If these deliberations lead to the application of manual electrode welding, TIG welding or partially mechanized gas-shielded metal-arc welding, the high wage costs and ancillary wage costs in Germany play a fundamental role in the consideration of the economic viability and frequently lead to the relocation of this production to low-wage countries.

The increased utilization of high-strength steels in structural steel engineering and the connected more stringent requirements on the precise compliance with the welding parameters are leading to greater mechanization and also to the utilization of laser welding in this sector.

Hand-guided welding processes are used regularly in Germany as well merely in special cases when the expense of transporting the parts is not in a reasonable ratio to the welding expense and for assembly welds.

Joining Technology in Installation and Mechanical Engineering. Statements similar to those in metal construction are also generally applicable to installation and mechanical engineering. Here as well, the manual processes with the different materials are replaced by mechanized and automated processes. In the future, the arc-welding processes used will be extended by laser welding and, in the case of greater material thicknesses, possibly by electron beam welding; this will also happen in the manufacture of large-diameter pipes. Current users are following developments in friction stir welding with great interest, possibly for the welding of steels as well.

Not only in the chemical industry but also in energy generation, process temperatures are rising constantly, thus leading to ever more stringent requirements on corrosion resistance. As a result, other materials such as nickel and cobalt alloys must be used instead of steels. The associated subsequent developments with regard to welding filler materials and welding processes must take place concomitantly with the development of materials for semifinished products.

Joining Technology in Shipbuilding. The changeover of welding processes in shipyards to laser-welding processes or to a combination of gas-shielded metal-arc and laser welding as a hybrid process with and without filler material is currently taking place in German shipbuilding. Although the costs of welding are not reduced very much (if at all) in this case, the benefits of the new process are to be found in the lower distortion caused by the drastically reduced weld volume and in the savings on expenses related to unnecessary straightening work. As an example, Fig. 7.213 shows laser-welded sandwich panels for shipbuilding.

Joining Technology in Aircraft Construction. Until now, riveting has been the basic joining technology in aircraft construction. For the first time, laser welding

Component
Sandwich panels for shipbuilding

Joining task
Connection of ship-deck panels (sandwich decks)

Special aspects
Smooth surface
Production of long welded joints → problem of welded-slot expansion

Material
Stronger steels and light-metal alloys

Welding techniques
Laser-beam welding with a 12 kW CO_2 laser

Advantages of laser-beam welding
Higher processing rates → high productivity
Low heat supply → low deformation → short finishing times
Improved strength and rigidity

Fig. 7.213 Trends in laser beam welding

is being applied in the wing area and in the fuselage. However, stiffeners are not the only thing that should be welded and not riveted. Developments for butt welding in the skin area have been initiated. Furthermore, adhesive-bonding technology is becoming increasingly significant in aircraft construction.

Joining Technology in the Product Life Cycle

As was already described, in various industries, the competitive situation of joining technology will, in the future, necessitate even stricter orientation to the product and to the product life cycle. The product life cycle begins with the idea for a product, with a customer order, or with a recognized market need. This leads to product development. The product life cycle ends with recycling or waste disposal. Between the beginning and the end of the cycle, the product is tested, manufactured, used, maintained, and, if necessary, repaired. In all these phases, joining technology must be taken into consideration from the very start. For automobile manufacturing, such strategies have already been described in detail. In the future, such strategies will, however, also have to be transferred to other areas of application of joining technology as well as to other sectors.

In this respect, it is necessary to pursue approaches not only relating to individual phases of the product service life but also, in particular, extending across different phases. Using the example of the automation capacity of the recycling of joined components, Fig. 7.214 shows an example of a phase-oriented approach.

Additional examples of such approaches that are related solely to one phase in each case are as follows:

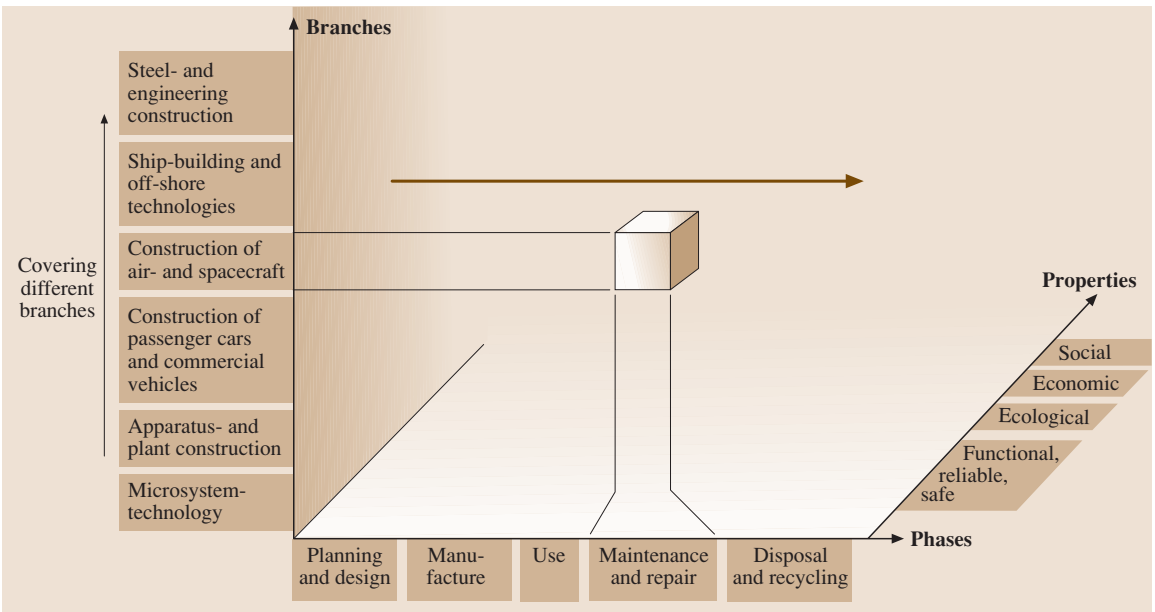


Fig. 7.214 Joining in product life cycle

- Designing in a way that allows for joining
- Real-time quality systems to control production by means of joining technology
- Long-term behavior of joints
- Economically viable processes for the repair and maintenance of joined components
- Economically viable processes for the dismantling of products
- Recycling concepts for multimaterial components

The interconnections in joining technology with respect to the design, composition, and manufacture of a product have already been elaborated in detail (cf. Fig. 7.214). Despite this level of knowledge about joining technology, there is still a need for further action and research. This is due to the fact that, until now, not all the aspects and influences that may arise during a product life cycle have been considered in good time and in an appropriate way. There is no all-embracing concept of joining technology that takes into account all the phases of the product life cycle and their interactions with each other (Fig. 7.215).

Approaches extending across different phases of the product life cycle are, for example:

- Designing in a way that is suitable for maintenance and repair
- Designing in a way that is suitable for recycling
- Cases of damage and repercussions on design and production

The universal simulation of joining processes and properties in all the phases of the product life cycle is an interesting example of such consideration extending across different phases of the product life cycle (Fig. 7.202).

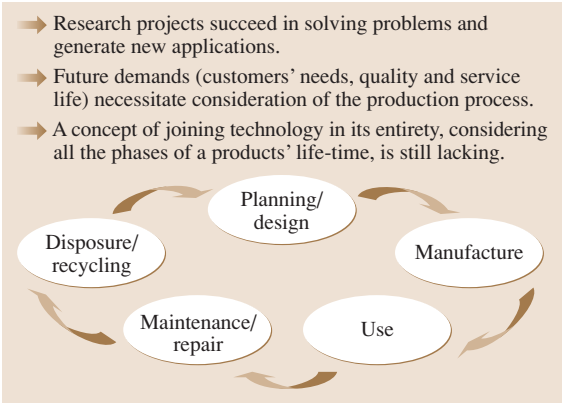


Fig. 7.215 Joining in product life cycle – initial situation

The consistent orientation of joining technology to the product life cycle may generate further value added and therefore has great economic significance for companies engaged joining technology. For the future of joining technology, it is thus necessary to strive for consider all the process chains from technical and economic viewpoints. In this respect, the integration of joining technology into automated production processes is playing an increasingly important role. The isolated consideration of individual joining processes or manufacturing steps must be given up in favor of the consideration of the manufacturing chain as a whole with the processes to be integrated and their interactions. The future will be characterized by the simplification of installation operations, improvements in sensor equipment, flexible and intelligent clamping and handling techniques, and integrated quality assurance – of course, supported by the universal use of simulation tools.

Summary and Outlook

For the technically and economically optimum utilization of joining processes, in the future, it will be essential to consider joining technology over the entire product life cycle. Interdisciplinary approaches and actions are imperative. In this respect, attention should focus on the system connections between materials, manufacturing, the joining technology, and the product. Lightweight construction will be a technological driving force behind joining technology in the future as well. Material developments with regard to the steels, aluminum alloys, and magnesium alloys will set new requirements on joining technology, and the multimaterial concept will gain acceptance. As a result, not only will increasing significance be attached to adhesive bonding and mechanical joining processes, but the thermal joining processes will also make enormous development advances in these application fields and will safeguard and extend applications.

The following trend statement may be made for the welding processes: gas-shielded arc welding retains a dominant position and the latest research supports the applications and further developments of these processes. The processes of resistance welding will become less significant despite considerable innovation efforts, and beam-welding processes (especially laser-welding processes) will become much more important due to the rapid transfer of the findings from research to applications. Combinations of welding processes (basically consisting of arc and laser welding) will belong to the state of the art in the future. New welding processes

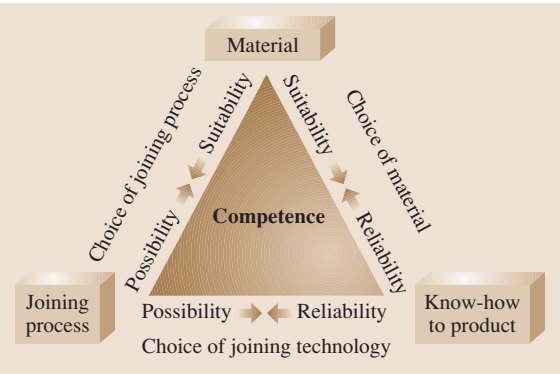


Fig. 7.216 Competence management in joining

such as friction stir welding will open up new applications.

The increasing complexity of joining technology is necessitating competence management in companies. Here, competence is defined as the joining together and use of relevant knowledge in materials engineering and in fabrication and joining technology for a very specific purpose. In this respect, the competence of a company is more than the sum of the competence of the individuals, but only when it is managed correctly. With reference to the joinability relationships already specified in Fig. 7.204, it is possible to develop competence management for joining (Fig. 7.216). The objective must be to assess and implement the experience available in a given company, the existing facilities, and proven processes in a proper and cost-effective way over the entire product life cycle. With attention focused on these conditions, joining technology will continue to account for a large proportion of the value added in the producing sector with prospects for further growth.

Joining technology will gain acceptance even in difficult economic conditions. The consistent further development of joining technology entails investment in appliances, processes, and infrastructure – however, particularly in people and in their knowledge and abilities.

7.4.2 Trends in Laser Beam Machining

By improving the laser beam source, new applications for laser beam welding could be found. With an increase in the maximum power output and improvements in the quality respectively the focusability of the beam, it has now become possible to perform joining on plate material as it is used in rail-vehicle-, plant- and apparatus-, or shipbuilding with the help of laser beam technology. Along with the enlarged scope of its usage, laser beam welding, as practiced in manufacturing engineering, its operation technology, and process-specific control engineering have all undergone innovation. But especially improvements in the laser systems have made the practice of laser beam welding simpler, safer, faster, and more reliable.

Laser Beam Welding

Basic Principles. Laser beam welding is a fusion welding process. Through the laser beam, the laser produces the energy necessary for welding. It is led to the focusing lens system by mirrors or optical fibers. The lens system then focuses the laser beam on the joint. Depending on the intensity generated there, different processes of beam–material interaction will take place. When the intensity is low, most of the radiation will be reflected by the workpiece, and only a very small part will be absorbed by the metal in a thin layer ($< 1\ \mu m$) at the surface and transformed into heat. The energy input into the workpiece is achieved by heat conduction. With increasing intensity, the workpiece is locally heated by the laser radiation that is absorbed. When the melting temperature is reached, a molten puddle forms as the time of influence is increased. If the parameters are chosen so as to maintain a stationary state, the process is described as thermal conduction welding (Fig. 7.217a).

If, however, the intensity is further increased so that more energy is absorbed than can be dissipated by heat conduction, then the enormous energy density of the laser beam in the focus will cause the metal to vaporize. The pressure of the metal vapor that flows off

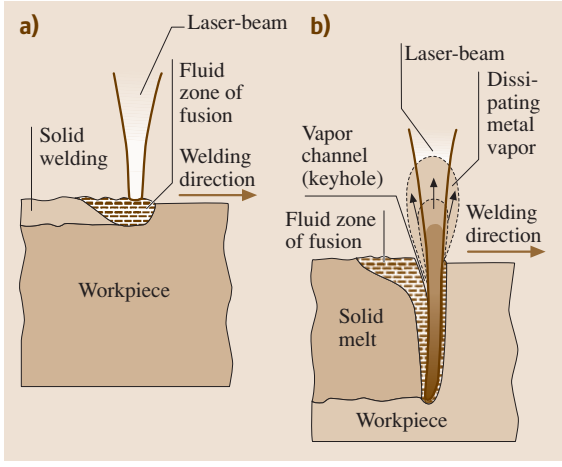


Fig. 7.217a,b Principle of laser beam welding: **(a)** Heat conduction welding, **(b)** deep welding

creates a steam passage in the melt, the so-called *keyhole*. Typically, the diameter of such a capillary steam tube shows the magnitude of the beam diameter (0.1 to 1 mm). Characteristic threshold intensities for capillary formation lie in the range of 10^6 W/cm^2.

In laser beam penetration welding, the system of capillary steam tube and surrounding molten bath is led along the assembly line. The molten bath flows around the capillary on both sides, comes together behind it, and, when solidifying, forms a joint (Fig. 7.217b) [7.157].

Trends in Systems Engineering. The motivation behind the development of lasers for industrial material processing has always been the demand for higher performance together with higher quality of beam. With the use of the traditional lamp-pumped Nd:YAG solid-state laser, increased performance of the laser is always connected to a decrease in beam quality, due to the thermal lens formed. By using laser diodes instead of pulsed-light sources, the thermal lens can be reduced; however, the principal decrease in beam quality will remain even if the laser performance is increased. The situation is different with a disc laser. This laser eliminates the thermal lens in the crystal so that a high performance together with a good quality of the beam can be obtained. With this, important new preliminaries for new laser applications are provided by achieving smaller focus diameter, longer distances between the optical machining system and the workpiece, and, last but not least, a higher focus depth of the focused laser beam [7.158].

In welding and cutting, the improved beam quality of the disc laser can directly result in a reduction of the focus diameter, which then causes the power density of the laser beam on the workpiece to increase to the square. The machining threshold moves toward lower performance, and the laser welding penetration effect begins sooner. The range of welding parameters is enlarged, and thus process reliability is increased. By the laser welding penetration effect that comes into play already at lower performances the weld seam width can be kept narrow. From this will result a low heat input and, thus, less distortion of the structural parts. Thus, an efficient and precise welding of thin sheets becomes possible. In addition, considerable penetration depth and welding speed are achieved [7.159].

Also when welding with scanner optical systems, the beam quality may be used to much advantage. Scanner optical systems with their galvo-mirrors deflect the laser beam within a working area. The focus diameter and working distance depend directly on the beam quality of the laser used. With the rod systems currently used and with their beam quality of 16 mm mrad, only working areas of approx. 150 mm side length could be accommodated when welding within a typical sheet thickness range of 1–2 mm. When working with a disc laser, an area of up to 300 mm side length can be covered without difficulty with the smaller scanner systems. This so-called remote welding procedure makes redundant the otherwise necessary system of axles with which the focusing optical system or the structural component is moved, but it is also of interest when combined with robots [7.160].

In the kilowatt range of performance, predominantly CO_2 lasers, longitudinally diode-pumped Yb:YAG lasers, transversally initiated Nd:YAG rod lasers, and Yb:YAG disc lasers are currently used. CO_2 lasers differ from other beams by almost diffraction-limited beams even at a performance of several kilowatts (typically $M^2 < 2$). With a wavelength of $\lambda = 10.6\,\mu$m, however, the beam parameter product of a CO_2 laser cannot be smaller than 3.4 mm mrad ($M^2 = 1$). Currently, in the kilowatt range, solid-state lasers do not yet produce diffraction-limited beams, but with 2.8 mm mrad ($M^2 = 8$) at 1 kW medium performance (rod laser) and 7 mm rad ($M^2 = 21$) at 4 kW (disc laser) they already offer excellent focusability with their short wavelength, which is ten times shorter than that of other lasers. In addition, diode-pumped solid-state lasers reach a high efficiency of greater than 20% (electrical-optical without cooling). Recently, fiber lasers have also appeared on the market that, by incoherent collimation of several fibers, offer a capacity of several kilowatts as, e.g., 1 kW power output with a beam parameter efficiency of 6 mm mrad ($M^2 = 18$) or 4 kW with 20 mm mrad ($M^2 = 59$) [7.161, 162].

The unique advantages of fiber lasers are their unsurpassed efficiency, excellent beam quality, low volume and weight, and outstanding robustness. These advantages can be attributed to their fiber-optical microstructure. Recent investigations have shown that their high performance can be achieved with several evanescent wave guides without impairing the beam quality. With such developments, fiber lasers are also capable of producing beams in the kilowatt range that offer an excellent focusability. The form of the modes can be determined by an appropriate design of the wave guides (distribution of the refraction index) [7.161].

It must be pointed out here that CO_2 lasers are available today not only with a 10.6 μm wavelength but also up to a performance of 6 kW, with a 9.3 μm wavelength.

The reduced wavelength leads to a 10% increase in focusability and improves the launching ratio [7.163].

Innovative Process Modifications. Recently, process quality in laser beam welding – a highly dynamic process – has been the topic of numerous experimental as well as theoretical investigations.

To increase the process stability in the welding of aluminum, a purposeful arrangement of two focused laser beams, the so-called TWINFOCUS technique, suggests itself. By welding with one or several lasers with two focused laser beams at one joint, a cavity geometry immune to deficiencies and imperfections can be achieved that at the same time guarantees an unimpeded streaming out of the metal vapor and with it pinhole-free seams. A multibeam technique can be carried out by splitting the beam as well as by combining individual beams. With CO_2 lasers, splitting of the beam is achieved either by a deflecting mirror with tilted surfaces or by a parabolic mirror with two dishes. With Nd:YAG lasers, splitting can be achieved either by using a double fiber and two independent beam sources or by placing a beam splitter into the beam path [7.166].

A double-focus technique in laser welding, i. e., the use of two spatially separated foci producing a common pool, is a new method which has advanced from laboratory tests to manufacturing practice (Figs. 7.218 and 7.219). By artificial dilatation of the steam passage a distinct increase in process stability in laser welding of aluminum has become possible so that weld pores and weld puddle eruptions can be almost completely avoided. However, dilatation of the steam passage will at the same time lead to a decrease in process efficiency [7.167].

Using laser beam sources with the highest beam quality it is possible to build up a focus matrix that allows the advantages of the double-focus technique, i. e., good quality of the weld seam, to be achieved with high process efficiency [7.164].

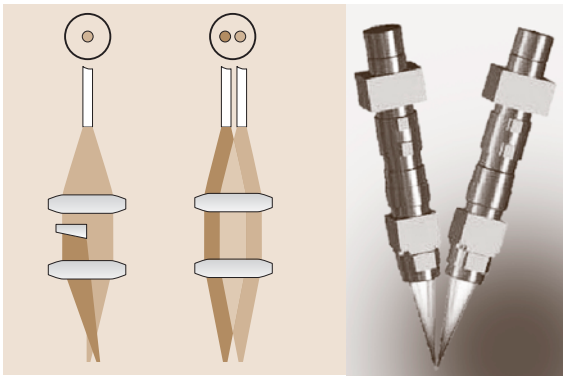


Fig. 7.218 Generation of double focus (after [7.164])

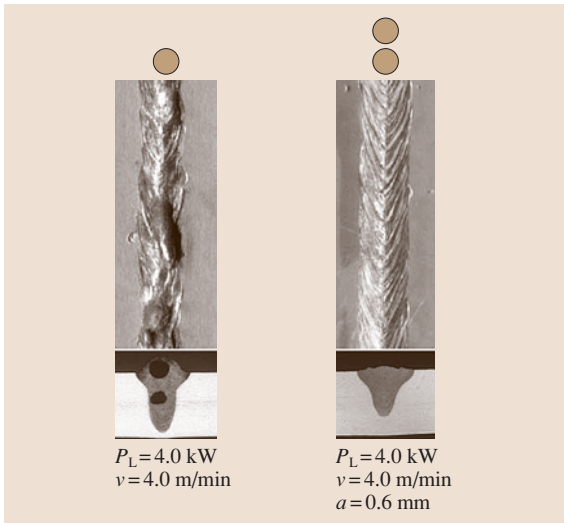


Fig. 7.219 Increase in quality by using double-focus technique

Another method for increasing process stability is magnetically supported laser beam welding, based on the application of an external magnetic field during welding. With this variant type of process, the flow behavior in the weld pool of molten metal can be changed by utilizing magneto-fluid-dynamic mechanisms in such a way that it will be stabilized and a higher welding speed will be arrived at. Considered in detail, the following results are verified: humping can be subdued, the final run quality is improved, spattering during weld-

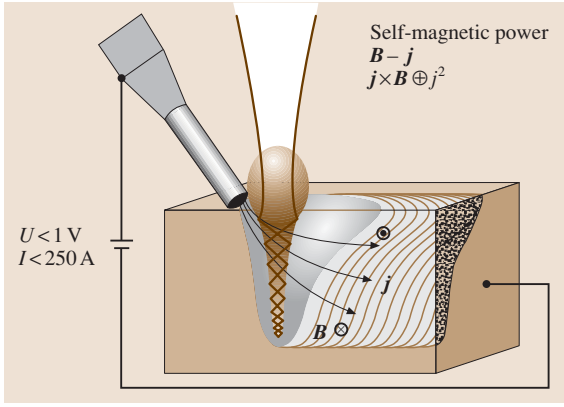


Fig. 7.220 Diagram of generation and utilization of self-magnetic power (after [7.165])

ing is reduced, fluctuations of the plasma flame are suppressed, and, last but not least, process stability is increased [7.168].

Every life wire is surrounded by its self-induced magnetic field, which is why energies are generated in it that act perpendicularly to the direction of both the electric current density j and the magnetic field B; their amounts are proportional to the square of electric current density. This principle is used in the following application (Fig. 7.220). An external current source provides direct current, which is applied either by a tungsten electrode moving just ahead of the molten pool or by a filler wire. The resulting self-magnetic power will influence the flow within the weld pool respectively the results of the welding operation. Both of the last mentioned methods (external magnetic field and external current) can be used together for influencing the process [7.165].

Traditionally, pulsed Nd:YAG lasers are typically used in the fields of boring and drilling, fine-edge blanking, and spot and microwelding. Due to their pulsed-energy release, pulsed solid-state lasers are extremely well suited for spot-welding operations. In seam welding, however, pulsed lasers are also frequently used in such areas that, due to their geometrical conditions, offer a challenge in the manipulation of a heat source, are demanding with respect to their metallurgy, or are problematic in their launching behavior. Beam penetration of pulsed solid-state lasers typically amounts to values from several tenths of 1 mm to about 4 mm. Mere deep welding with a high aspect ratio is usually not found with pulsed solid-state lasers [7.169].

Another forward-looking technology that expands the range of applications of pulsed Nd:YAG lasers is the so-called SHADOW method. With this method, joints several millimeters or centimeters in length cm are welded with only a single laser pulse by high feed speeds which are attained using fast rotating axles or scanners. Typical pulse lengths are around 10 to 50 ms for welded seams with possible lengths of up to 80 mm. The efficiency of the SHADOW method results from the special manner in which energy is input into a material. While for conventional seam welding with a pulsed Nd:YAG laser a high proportion of heat conduction is characteristic, the SHADOW method typically has a welding speed higher than the heat conduction in the material being welded [7.170].

For labeling, scanners and laser beams have for a long time offered a perfect combination, though rather for smaller-scale projects. Deflection of the laser beam via two extremely fast rotatable mirrors might now also revolutionize welding (Fig. 7.221). A decisive advantage of this principle is that positioning times are reduced to practically zero, which drastically diminishes processing time. The decisive advantage of laser welding with a scanner compared to traditional laser welding consists in a considerably faster positioning of the laser beam. When welding structural parts with several welds in a conventional way, there will often be longer traverses with corresponding nonproductive times, whereas these times move toward zero when using a scanner. The high positioning speed and flexibility of welding with a scanner offers additional advantages: heat distortion can be reduced as welding operations distributed over the workpiece can be executed without delay.

By combining the advantages of scanner-based processes and robot-guided machining, parts can be processed extremely flexibly and over longer distances (Fig. 7.222). Of special importance are optical sensors that are able to recognize the position of the weld and automatically control the process. The lasers used must meet extreme demands as to the beam quality, and, considering the very long focal distances, they must reach sufficiently small focus diameters. Robot-guided scanner welding is mainly used in automobile manufacturing.

Currently, the actual operating and production time with laser beam welding is often short, compared to nonproductive times, which mainly result from operating the laser heads. With remote welding, however, a very fast sequential welding of spot or short welds

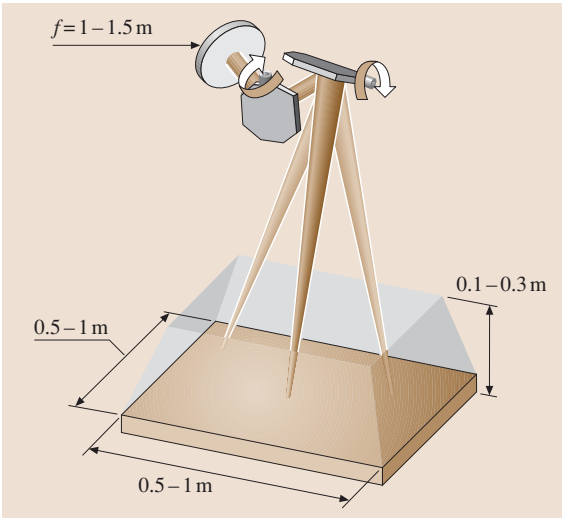


Fig. 7.221 Principle of scanner welding (source: Trumpf)

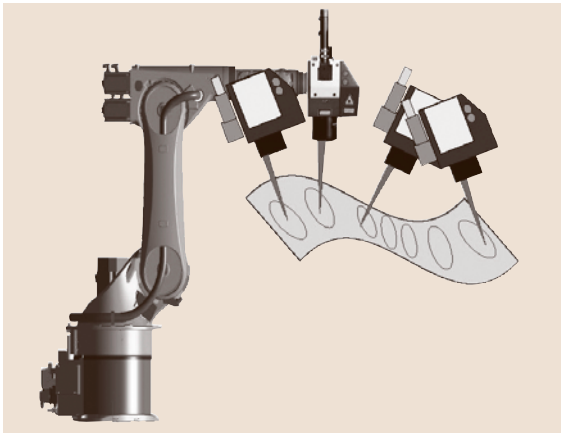


Fig. 7.222 Scanner welding with robot

can be accomplished with positioning speeds of roughly 700 m/min. This is advantageous, above all, in welding structural parts with numerous short welds distributed over the workpiece, and it allows an actual laser operating time of above 90%.

The remote method might mean a quantum leap for production. The process promises the substitution of techniques used so far as well as new types of products. In car body manufacturing, remote welding reduces investment costs by 30% and the cycle time by 60% compared to spot welding. Correspondingly, a study on the remote method for the year 2015 predicts its proportion of 8–10%, referred to the mass of joining methods.

Trends in Quality Monitoring. As the process of laser beam welding is only restrictedly fault-tolerant, it follows that the required quality of the weld might not be attained. Extremely high quality demands as required, e.g., in welding car bodies make an online assessment of processing quality absolutely necessary.

Typical signals from emissions that closely correlate with processing quality come from the heated respectively molten material (thermionic emissions in the IR wavelength range), from hot metal vapor as plasma radiation (UV wavelength range), or as metal halide lamp (visible wavelength spectrum). Upon reaching the evaporating point of the material (quantitative threshold intensity typically several 10^6 W/cm) the so-called deep-welding effect occurs. Here, the flowing-off metal vapor reaches an equilibrium of forces with surface tension and hydrostatic respectively hydrodynamic pressures so that a vapor cavity is formed in which, due to multiple reflections at the cavity walls, laser radiation will be absorbed. The unabsorbed radiation will be reflected backwards out of the cavity. Here, information may be acquired on the quality of the weld deep from the inner part of the material. Methods for controlling the laser welding process are mainly based on the above-mentioned process emissions [7.171].

The most reliable statement on quality can be made by combining different measurement methods. In this respect, plasma diagnostics, monitoring of the molten pool, and, by the increasing use of the Nd:YAG solid-state laser in industrial production, the laser power reflected from the workpiece are primarily used as measured parameters for online process control [7.157, 172].

Laser Beam Cutting

In this process, the focused laser beam hits the workpiece where it locally fuses the material and also partly or completely vaporizes it. By the impulse of a gas jet emerging from a nozzle the material is removed and leaves the kerf due to the relative movement of beam and workpiece (Fig. 7.223). The gas jet is, at the same time, meant to protect the focusing optical system against vapor and weld spatter.

Of all the techniques of laser processing of materials, laser beam cutting is the most widely used. This process became widely used as a result of its high flexibility, the great variety of manufacturing processes, and the excellent quality of the cut were of crucial importance. With this technique almost any material can be cut, the suitability for laser beam cutting depending on the absorption behavior of the surface, the ignition, melting, and evaporation temperatures of the material, and its thermal conductivity.

In laser cutting, three versions of the process are to be distinguished:

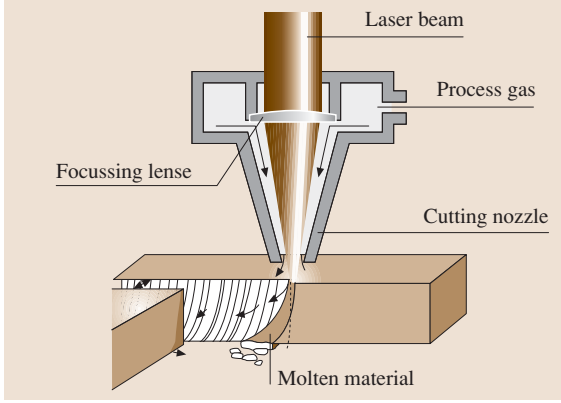


Fig. 7.223 Principle of laser beam cutting

- Oxygen laser cutting
- Fusion laser cutting
- Evaporative laser cutting

Oxygen Laser Cutting. The laser beam heats the material to ignition temperature. The oxygen injected into the kerf burns the material and expels the slag formed. The combustion process generates additional energy. With the quality of the cut being continuously high, a distinct connection between the purity of the oxygen and the maximum possible cutting speed can be proven.

Fusion Laser Cutting. In this version of the procedure, the material gets fused in the crossover point by laser radiation. The melt is expelled from the kerf by an inert gas. High-pressure fusion laser cutting is proving to be increasingly successful in oxide-free cutting of stainless steels. It is also successfully used in cutting mild steels and aluminum. As a rule, nitrogen is used as the cutting gas. The cutting gas pressure at the cutting nozzle can be 20 bar and above.

Evaporative Laser Cutting. In evaporative laser cutting, the material to be cut is evaporated at the crossover point of the laser beam. An inert gas, e.g., nitrogen or argon, expels the byproducts from the kerf. This cutting process is used with materials that have no liquid phase or melt, as is the case with paper, wood, several synthetic materials or plastics, textiles, and ceramics.

At present, CO_2 lasers with performances of up to 5 kW and Nd:YAG lasers with performances of up to 2 kW are in use for laser cutting. Special CO_2 cutting lasers with performances of up to 5000 W allow process-reliable machining of mild steel plates with a thickness of up to 25 mm. With high-speed thin-sheet cutting, cutting speeds of up to 40 mm/min are achieved. New drive mechanics allows positioning speeds of up to 300 m/min [7.173].

So far, CO_2 lasers have proven suitable tools for fast 2-D laser cutting of thin sheets due to their good focusability and high laser beam performances. By increasing the beam quality of solid-state lasers through the use of diode-pumped Nd:YAG lasers, with new resonator programs, launching into ever smaller fibers becomes possible with which, in the meantime, suitable focusing for high-speed cutting has become practicable.

So, for example, launching of laser beam performances of up to 4 kW in fibers of 300 µm diameter is feasible. The aim of the current development of solid-state lasers in the multikilowatt range is an increase in beam quality to about 5 mm mrad. At this level, the beam could be launched into an optical wave guide of 100–200 µm core diameter, and a laser would be available for machining that could be very finely focused as is possible with the best CO_2 lasers already in use [7.174].

Laser cutting has, until now, been dominated by flat-sheet cutting. This is mainly due to four factors: the dynamics and precision of five-axle laser cutting systems have been too poor, and cutting speeds have been much lower than with cutting flat sheets. Programming of five-axle cutting systems is very time consuming and inaccurate. The price of the system is higher than that of two-axle cutting systems. Finally, the market for parts cut in three dimensions is much smaller than that for flat-sheet cutting [7.175].

The use of laser cutting systems becomes especially critical when cutting thick plates. A 4 kW CO_2 cutting laser can, e.g., facilitate processing of mild steel plates up to 25 mm thick, using oxygen as the cutting gas. In conventional laser cutting, an increase in the range of cuttable plate thickness can only be reached by further increasing the laser beam performance available. However, with the newly developed LASOX technique (laser-assisted oxygen cutting) it is, given a considerably lower laser beam performance, possible to cut plates more than 50 mm thick with a laser beam performance of 2 kW. This is achieved by modifying process conditions at the front of the kerf. The laser beam is defocused and will heat the workpiece surface only to ignition temperature without melting it. The diameter of the laser beam on the plate surface is larger than the diameter of the cutting oxygen jet. Under these conditions, the cutting process is similar to oxy-gas flame cutting rather than to the conventional laser cutting [7.176, 177].

The increasing availability of high-power lasers above 6 kW of good mode quality, TEM 01 or better, for fusion laser cutting opens up new possibilities for entrepreneurial reorientation. According to laboratory tests, a cut thickness of more than 30 mm is feasible with a laser beam performance of 9 or 12 kW. At present, the perspective limit is considered to be a material thickness of 45 mm [7.178].

In fusion laser cutting in the plate range, the incomplete expulsion of the melt from the joint presents considerable difficulties. By applying liquids instead of gases this problem may be eliminated. In comparison to gas jets, water jets produce a higher impulse response. In this way the material is more effectively removed from the kerf [7.179].

Already for a considerable time, laser users have been observing a connection between cutting quality and the quality of the steels to be cut. For this reason, special steels have been developed that meet the demands of the respective standards for sheet metals as well as optimal cutting results [7.180].

An innovative microjet procedure (LMJ) is based on a fine laminar jet of water 40–100 μm in diameter into which the pulsed laser energy is introduced. During the pulse time the material will melt and at the same time be removed by the water jet. During the pulse interval, the kerf is cooled by the water jet. Thus, heat input into the workpiece is distinctly reduced. The process is especially suitable for use with temperature-sensitive materials such as, e.g., silicon wafers or ceramic materials where the brittleness of the material or its extreme hardness make machining with other methods more difficult [7.182, 183].

Laser Beam Drilling

Drilling with a laser beam was among the first industrial applications of material processing by laser. Already in the 1960s, sapphire and ruby bearings for clockworks were laser-drilled in large numbers in the Swiss clock- and watch-making industry [7.184]. The early 1970s saw the introduction of the industrial use of laser beams for drilling metal.

Laser drilling is a thermal abrading process. The power density necessary for abrading is reached by focusing the laser radiation onto the workpiece with the help of a lens system. In general, drilling is based on the material getting melted within the area treated by the laser beam and being vaporized. Whether the material is abraded in the form of melt or vapor or in both phases depends on the material's properties and intensity of radiation. After the launching phase, the evaporation front moves into the workpiece [7.185, 186]. A gas jet directed coaxially at the laser beam supports the expulsion of the melt out of the created hole and simultaneously protects the focusing lens system against spattering metal. As for the process technology, the following methods are used in drilling: single-pulse, percussion, trepanning, and twist drilling (Fig. 7.224).

In *single-pulse drilling*, the workpiece is drilled through by a single laser pulse. This procedure is used for drillings with small diameter in thin sheets with an aspect ratio of $< 1 : 15$.

In *percussion drilling*, several laser pulses are directed at the same place on the material. This is necessary when a deeper hole is drilled than can be done with single-pulse drilling. Thus, holes several millimeters deep can be made. With this method, high-quality drill holes with an aspect ratio of up to $1 : 200$ can be achieved. Percussion drilling also offers the advantage of a better repeatability of drill hole geometry compared to single-pulse drilling.

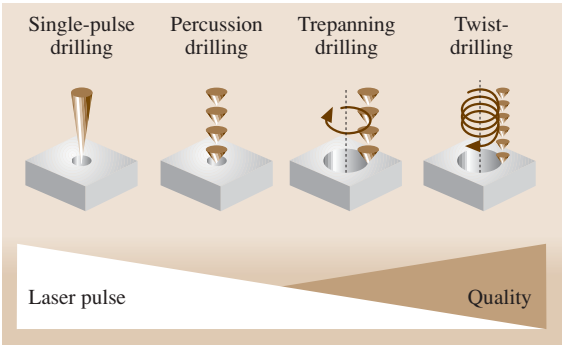


Fig. 7.224 Process technologies in laser beam drilling (after [7.181])

Trepanning drilling, also known as the *circular cutting method*, is used for producing larger holes. With this method, the dimension and precision of the drill hole diameter is determined by the relative motion between the workpiece and the laser beam. This technology roughly corresponds to laser cutting.

An additional increase in precision by further decomposing the drilling process in ever more individual steps becomes possible by the so-called *twist drilling*, developed from trepanning drilling. Various scientific investigations have proven that a drill-hole quality that has never been achieved in ceramics and steel is possible with this method [7.181].

The centerpiece of the trepanning optical system consists of three specially designed beam splitters for aimed deflection of the laser beam. During the process, all three beam splitters rotate around the fiber optical axis. This allows one to adjust a desired phase shift proportional to the radius of the helix on the material. Integrating the trepanning optical system into the polarization adjustment will further improve quality. Drag lines will be avoided, the outlet cross-sectional area will be circular, and process efficiency will increase due to higher absorption [7.187].

With all four processes mentioned above, drilled holes can be made perpendicular as well as angled to the workpiece surface. Blind holes can only be drilled into the material with single-pulse and percussion drilling [7.188, 189].

For drilling, mainly pulsed lasers are used. Many different investigations have shown in recent years that,

with shorter pulses, processing quality can be improved with respect to repeatability. This is essentially due to modifying the mechanism of removal with different pulse lengths [7.181, 190].

When drilling with solid-state lasers in the nanosecond range, holes with a melting film thickness of less than 1 µm can already be produced with suitable process technology. Further diminishing of pulse lengths into the range of picoseconds and femtoseconds is often propagated as a possibility for completely avoiding the formation of melt. Theoretical investigations have shown, however, that with metals a reduction of pulse length is sensible only down to about 10 ps (Fig. 7.225). An explanation for this is to be seen in the specific heating behavior of the crystal lattice on the ultrashort time scale [7.191]. This result, gained through an understanding of the process, should be taken into consideration by development engineers of beam sources.

Laser drilling is successfully used in many domains of engineering and industry. Typical examples of use in the motor manufacturing industry are, e.g., the production of blind hole grooves, structuring of surfaces with oil pockets for minimizing friction, and production of holes in diesel injection nozzles and in metallic filter discs. For example, with this procedure 100 holes/s can be produced in a metallic filter screen with a wall thickness of 0.95 mm [7.192, 193]. If there are special demands for precision holes with respect to cylindricity or surface finish, holes made by a laser beam can be refinished with conventional drilling methods. For these purposes, spark-erosive and mechanical drilling methods are used first. In these cases, at first a somewhat smaller pilot hole is made by a laser beam and is finished to size using a conventional drilling method [7.194].

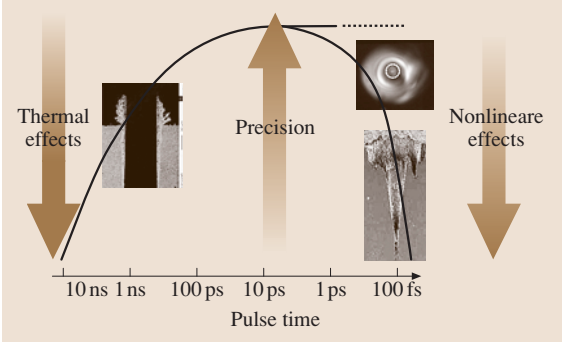


Fig. 7.225 Precision dependent from pulse length in drilling metals (after [7.191])

7.4.3 Electron Beam

Introduction

The principle of electron beam (EB) technology is based on high-voltage-accelerated electrons that can be used as a tool for material treatment, especially welding. Nowadays, EB welding is firmly established in many areas of manufacturing and is generally accepted for its reliability and efficiency. Joining tasks can be used in a variety of tasks ranging from foil welding with thin plate thicknesses of just a few tenths of a millimenter up to thick-plate welding with achievable weld depths of more than 150 mm in steel materials and more than 300 mm in aluminum materials. Moreover, almost all electrically conductive materials are weldable, and many of those materials may also be joined in material combinations. The high power density in a range of up to 10^7 W/cm^2, which is typical of EB welding, and a connected depth-to-width ratio of weld (up to 50 : 1) allows for a large variety of possible applications of this joining process. As the electron beam is electromagnetically deflectable and quasimassless, progress in the field of control technique and increased processor performances has even extended the multitude of applications, and there are more in the offing.

Apart from the EB variations where welding is carried out with high or low vacuum in a vacuum chamber, there is also the possibility of applying the EB in the atmosphere (nonvacuum) as a joining tool.

Basics of the Process

Electron Beam Generation and Guiding.

Generation of the Electron Beam. In modern electron welding machines, triode systems for the beam generation are generally applied. These systems are composed of anode, cathode, and controlelectrode (Wehnelt cylinder). The electrons that are necessary for beam generation are emitted from the cathode by thermionic emission. The cathode is made of a material in which the work function that has to be performed by the electron in order to leave the material is comparatively low. The cathode material must show a high electron emission rate and also be high-temperature resistant. The material should, moreover, also guarantee a relatively long cathode life. Appropriate materials are tungsten and tantalum. Mostly direct heated cathodes are used, which are flowed by current and heated thus by Joule resistance heating. By the application of an electric high-voltage field, the electrons are supplied with kinetic energy in order to emit them from the electron cloud and to subsequently accelerate them. Depending

on the height of the applied voltage, the electrons may be accelerated up to two thirds of the speed of light. For the generation of the electric field, between cathode and anode (the anode is arranged across from the cathode), the acceleration voltage is activated [7.195]. By the application of a control voltage between the cathode and a control electrode a barrier field is generated in this triode system that forces the emitted electrons back to the cathode. Thus the beam current is controlled by alterations of the control voltage as by the decrease of the control voltage more electrons pass the barrier field toward the anode. Due to its particular shape, which can be compared to a concave mirror used as in light optics, the controlelectrode affects the electrostatic focusing of the electron beam. After having passed the anode, the electrons have achieved their final speed, and the EB is focused and deflected by means of electromagnetic focusing lenses. The focusing effect leads to the constriction of the EB, the so-called crossover.

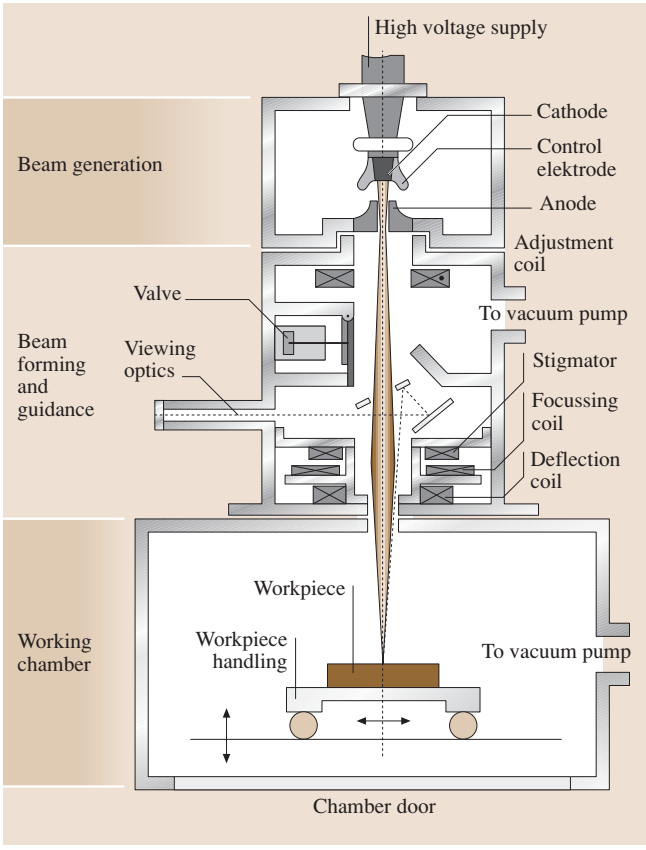


Fig. 7.226 Diagrammatic representation of an electron beam welding machine

Beam Manipulation. The EB that diverges slightly after having passed the pierced anode is focused to a spot diameter of 0.1 and 1.0 mm by the following beam manipulation system in order to reach the necessary power density of 10^6–10^7 W/cm2. First, the beam is guided through the alignment coil onto the optical axis of the focusing objectives. One or several electromagnetic lenses bundle the beam onto the workpiece inside the vacuum chamber. Deflection coils that are positioned at various parts of the EB generator assist in the deflection or oscillating motions of the EB. A diagrammic representation of an EB welding machine is depicted in Fig. 7.226.

Deep-penetration Effect. When the electrons strike the surface of the workpiece, their kinetic energy is converted into thermal energy. Although the electron mass is very low, approx. 9.1×10^{-28} g, they have a high electric voltage potential that, at an accelerating voltage of 150 kV, allows electron acceleration up to a speed of $\approx 2 \times 10^8$ m/s. Not all beam electrons penetrate into the workpiece and release thus their energy to the material. A part of the striking electrons is emitted in the form of backscattered electrons, thermal radiation, secondary electrons, and X-ray radiation (Fig. 7.227).

As the electrons, owing to their low mass, that penetrate into the material achieve only very shallow penetration depths (up to 150 µm), a special effect is needed to obtain large weld depths, the so-called deep-penetration effect. The material is melted and vaporized in the center of the beam. This happens so fast that the heat dissipation into the cold material shows practically no significant effect. The resulting vapor is superheated

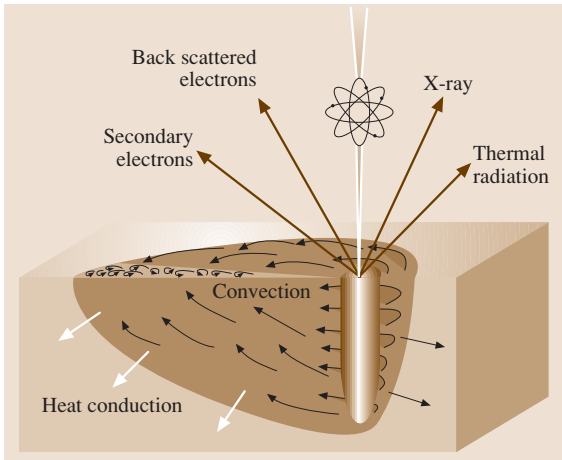


Fig. 7.227 Energy transformation into the workpiece

and extends at temperatures above ≈ 2700 K. The vapor pressure is sufficiently high to press the molten metal upwards and to the sides. A cavity develops where the electron contacts the yet unvaporized metal and heats this further. This leads to a vapor cavity that at its core consists of superheated vapor and is surrounded by a shell of fluid metal. This effect is maintained as long as the pressure from the developing vapor cavity and the surface tension of the molten pool are in equilibrium. The diameter of the vapor cavity corresponds approximately with the EB diameter. With a sufficiently high energy supply, the developing cavity penetrates through the entire workpiece [7.196]. The relative motion between workpiece and EB causes the material that has been molten at the front of the EB to flow around the cavity and to solidify at the backside.

Machine Components. An EB welding machine is composed of a multitude of individual components. The basic component of the machine is the EB generator where the EB is generated in a high vacuum, influenced by electromagnetic deflection coils, and then focused onto the workpiece into the vacuum chamber. As the electron beam in free atmosphere diverges strongly by the collision with air molecules and thus loses power density, welding is generally carried out in fine or high vacuum inside a vacuum chamber. For the vacuum generation in the beam generator and in the working chamber, different vacuum pumps are used. While in the beam generator a high vacuum ($p < 10^{-5}$ mbar) for insulation and for oxidation circumvention of the cathode is indispensable, the possible working pressures in the vacuum chamber varies between high vacuum ($p < 10^{-4}$ mbar) and atmospheric pressure.

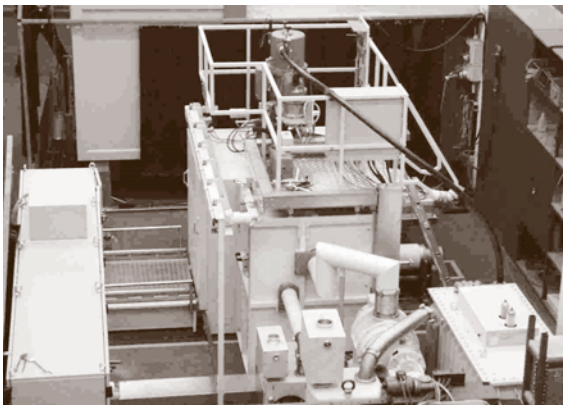


Fig. 7.228 Electron beam welding machine and peripherals

A shutoff valve, positioned between the EB generator and the working chamber, allows for the maintenance of the beam generator vacuum even when the working chamber is flooded. Besides the already mentioned modules, the high-voltage supply, the controls for the high-voltage supply, the vacuum pumps, the NC of the work table, and the operator areas are also necessary. The equipment is controlled via the operator console where all relevant process parameters are set and monitored. In modern equipment, parameter selection and control may be assumed externally by a computer and corresponding software. For the determination of optimum welding parameters for process control and also for the adjustment of the EB on the workpiece, viewing optic systems are necessary. One can use either simple light-optical viewing optics with a telescope or a camera system and monitor that partially represent the magnified section or an electron-optical system. In these systems the EB scans the workpiece surface at a very low power without melting it. The backscattered secondary electrons show, as in SEM, an image of the workpiece surface.

A disadvantageous effect, along with the increasing acceleration voltage, has, however, an exponentially increased X-ray radiation and also an increased sensitivity to flashover voltages. In industrial production a distinction is made between high-voltage machines with acceleration voltages of between 120 and 180 kV and low-voltage machines with acceleration voltages of max. 60 kV. Beam powers of up to 200 kW are used.

Potentials of Fast Beam Control. As the EB is an almost massless welding tool that is deflectable, noncontacting, and almost inertia free, it is possible to oscillate the beam with extremely high frequency. With this technique the EB skips between several positions with a frequency so high that the metallurgical influence on the structure is carried out at different points

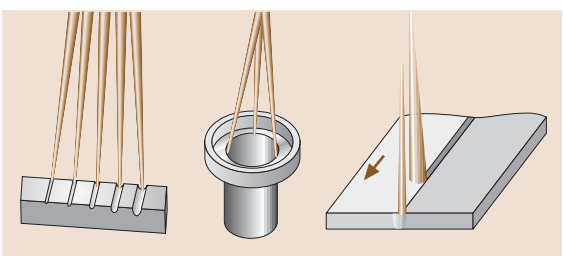


Fig. 7.229 Multibeam technique

simultaneously – due to the thermal inertia of the structure. It is possible nowadays to control the beam in such way that a simultaneous processing of the material due to Fig. 7.229 can be handled.

Production Machine Concepts

Apart from the further development of beam generators, the adaptation of the equipment to various demands is of considerable importance for the industrial application of EB welding. Using adapted vacuum chambers the evacuation times may, for many applications, be reduced in such a way that the necessary welding downtimes are not the decisive criterion against the use of EB technology. Different working chamber systems are now available to equip EB welding machines.

Chamber Machines. The most flexible variant is the universal working chamber where the workpiece is moved in two or three directions.

As an alternative to an NC coordinate table, revolving devices with horizontal or vertical axes of rotation are also applied. Typical chamber sizes reach from 0.1 to 20 m^3; machines with a chamber volume of up to 3500 m^3 are, however, also in use.

This vacuum chamber concept, however, entails comparatively high downtimes as the working steps *clamping the tool*, *entering the recipient*, *evacuation*, *welding*, *airing*, and *workpiece release* must be carried out one after the other.

Double Chamber and Lock Chamber Machines. Double chamber machines have two working chambers that are placed side by side. Either the beam generator is moved between those two chambers or the beam is deflected to one chamber at a time. Thus welding may be carried out in one chamber while the other chamber is loaded or discharged as well as evacuated. Figure 7.230, left, shows one of the variations of a double-chamber machine.

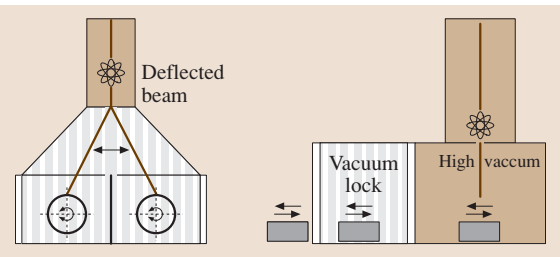


Fig. 7.230 Double chamber and lock chamber machine

Lock welding machines represent another equipment concept, Fig. 7.230, right. A high vacuum is permanently maintained in the chamber where the welding is carried out. Manipulation devices pass the workpieces through one or two prechambers. The machines have a position for loading and discharging, a lock for airing and deaerating, and a welding lock.

Cycle System Machines and Conveyor Machines. Cycle system machines are the right choice for welding similar or equal parts with equal weld geometries and axial welds. Underneath the chamber (generally small-volume), which accordingly demands only short evacuation times, a rotating jig with vertical, horizontal, or swivelling axes is fixed and is equipped with one or several loading stations. Thus loading and discharging as well as welding may be carried out at the same time. At new cycle system machines there is a low vacuum in the area where the jig rotates all the time. This leads to shorter evacutaion times from free atmosphere to low vacuum at the loading position and from low vacuum to high vacuum at the welding positions.

Especially for the manufacture of saw bands, a type of equipment that is the most productive but, on the other hand, the most inflexible, has prevailed on the market – the conveyor machines. These machines have the same operating principles as the lock welding machines where the workpiece is continuously transported over centering lips through pressure locks into the working chamber and from there again through a pressure lock. The inevitable leakage must be compensated by the vacuum technique.

Nonvacuum Electron Beam Welding

Already in the early stages of EB technology development, the method to guide the beam from the vacuum environment of the beam generator to the atmosphere was researched in Germany. This was the basis of the nonvacuum electron beam welding (NV-EBW) process. The substantial weld depths that characterize vacuum EB welding are not achievable with the NV-EBW method – those weld depths characterize the vacuum EB and are a result of its power density. The strong points of NV-EBW lie mainly in high-speed production. The achievable welding speeds reach up to 60 m/min when welding aluminum sheets and up to 25 m/min when welding steel plates. For a better energy coupling to the workpiece, to this day beam generators with an accelerating voltage of 175 kV are used [7.197, 198].

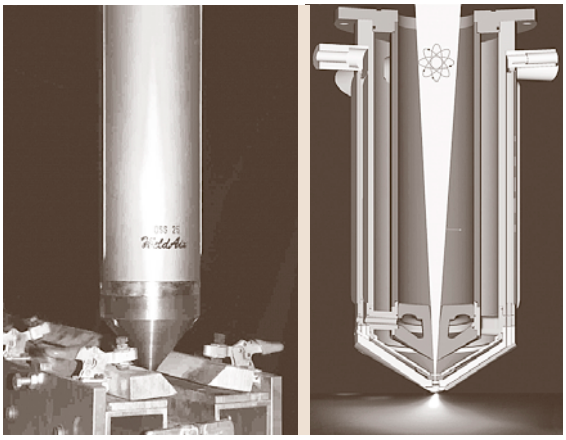


Fig. 7.231 Photo and principle of the ISF-non-vacuum nozzle system

Technology of NV-EBW. A solution to the vacuum-related restrictions is, following the modern state of the technique, achieved by guiding the vacuum-generated electron beam after it exits the beam generator to the atmosphere over a multistage orifice assembly and nozzle system. The pressure chambers that with a correspondingly higher pressure (10^{-2} and 10^{0} mbar) are connected after the beam generator chamber (vacuum 10^{-4} mbar) are evacuated separately and are separated, in vacuum terms, from each other by pressure nozzles. The EB is focused on the exit nozzle, which shows an inner diameter of 1–2 mm, Fig. 7.231. After its exit to the atmosphere the EB collides with air molecules and expands [7.196, 199].

The scattering of the EB is reduced by the increase in the accelerating voltage and by the application of helium as a working gas. For the effective utilization of the helium gas flow, a coaxial gas jet is applied at the beam exit outlet. The effect of the electron scattering at the gas molecules is additionally attenuated by the strong heating of the gas in the electron path. The gas density is reduced by that and the scattering also decreases [7.196–199].

After their exit from the orifice assembly, the electrons of the focused beam impinge on the material surface at a high speed and transmit their kinetic energy to the material lattice. As in vacuum EB welding the X-ray radiation must be shielded with sufficient radiation protection. In addition, the ionization of the air causes the generation of ozone, which must be neutralized. A radiation-proof working chamber may be designed with different materials and in optional dimensions. Limits to the component size are, therefore, not much higher in NV-EBW than in Nd:YAG laser welding.

Applications of NV-EBW. A high-power and out-of-vacuum EB is the ideal tool for welding conventionally manufactured sheets and sheet metal parts. The upper bead of the weld is similar to that of an arc weld and is thus not comparable to the typically narrow deep geometry of vacuum electron beam-welded joints. The method is characterized by a high energy efficiency, and its high available beam power yields a high power density even when the beam is expanded and allows high welding speeds (Fig. 7.232).

A classical application field of NV-EBW is the welding of components where several plates that form a flange weld are joined (Fig. 7.233). NV-EBW is perfectly suitable for this application. The broad beam fuses several plate edges simultaneously, which leads to a gas-tight and even joint. The flange weld and the lap joint are particularly suited for components where, after a very rough weld preparation, the desired result is a gas- and liquid-tight weld.

The materials that have been tested up to now with the NV-EBW method were uncoated and coated steels, light metals such as aluminum and magnesium and nonferrous metals such as brass or copper. Material combinations such as, e.g., the combination of steel and copper may also be realized with results comparable to vacuum EBW, without, however, achieving the higher weld depths of vacuum EBW. Supplementary, application tests on the use of filler wire in NV-EBW have been carried out.

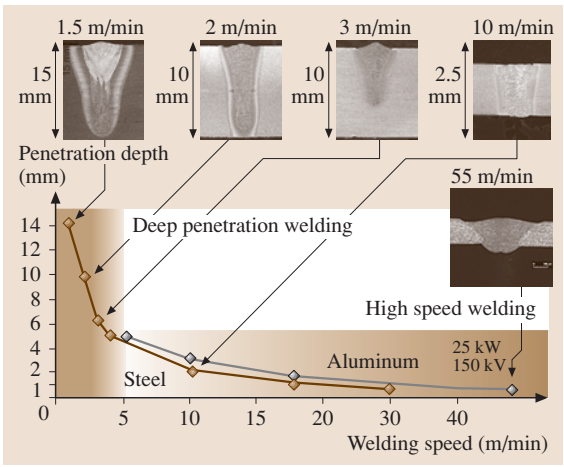


Fig. 7.232 Working range of NV-EBW with an acceleration voltage of 150 kV

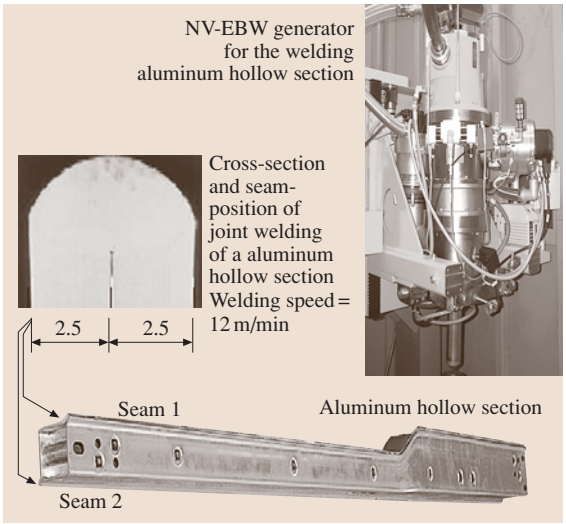


Fig. 7.233 NV-EBW equipment for the manufacturing of Al-hollow sections

Quality Assurance

Beam Measurement. For full exploitation of the advantages of the EB as a tool for welding, knowledge about all beam properties is necessary. The processes that occur in EB welding are very complex and are characterized by a multitude of different parameters such as, e.g., accelerating voltage, beam current, focus position, or power density distribution. For the determination of the various beam parameters and for the facilitation of the parameter transfer between different EB units a number of beam-diagnostic systems that apply different measurement principles are currently being developed [7.200–202]. The DIABEAM [7.203–205] has been developed at the ISF Welding Institute, RWTH-Aachen University. This system may be employed in almost all existing EB units and allows signal acquisition of the electron beam, up to a power of 100 kW. The DIABEAM measurement system was developed for easy determination of the focus position with slit measuring or a rotating wire, eliminating the need for complex and cost-intensive welding tests. The measurement and the three-dimensional display of the power density distribution across the beam diameter can be made by means of the apertured diaphragm. The other purpose of beam diagnostic is the prevention of negative influence, which may be caused by cathode adjustment, cathode distortion, variation of the vacuum level, etc., on the welding result by in-time identification of variations of beam characteristics. As a result of the three measuring processes (hole, slot, and rotating wire), the DIABEAM beam diagnosis system is suitable for a broad range of applications, especially for analyzing and assuring beam quality.

Sensor Systems. Scanning can, alternatively, be carried out via a slit or apertured diaphragm or via a rotating tungsten wire. The point of slit measurement with slit widths of 20 µm is the comparatively fast and simple determination of a signal that is in proportion to the beam intensity and, also, the determination of the beam diameter. The principle of slit-measurement with the appropriate voltage signal over the beam cross section is depicted in Fig. 7.234a.

With the slit measuring process, the core and edge areas of the beam can be compared by means of five different selectable diameters. Measurement of the beam diameter under varying working distances enables the caustic curve of the beam to be determined and displayed (Fig. 7.234b). In this way, the beam aperture can be determined precisely, thereby considerably simplifying the welder's selection of electrical and geometric parameters.

The application of a double-slit sensor enables online measurement of the deflection speed, which in-

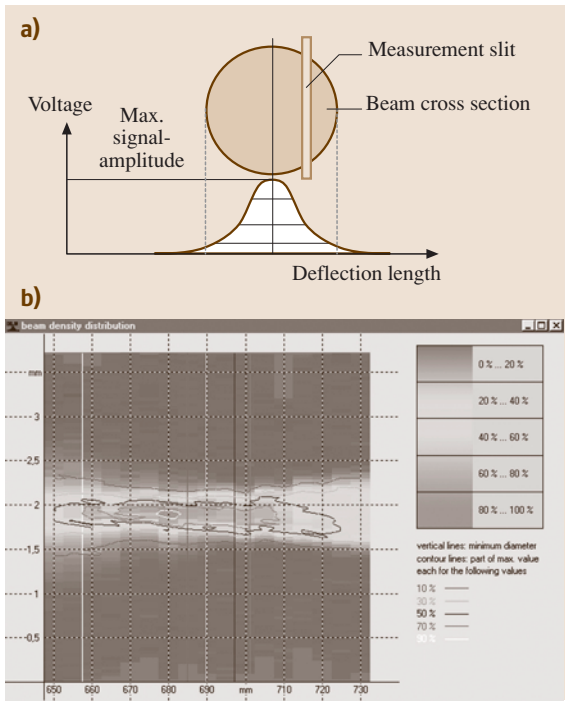


Fig. 7.234 **(a)** Principle of slit measurements, **(b)** results of a series of slit measurements

creases the precision of the measurement. Here, the beam is, at the start of the measurement, deflected from its neutral position transversely over both slit sensors. No deflection into the lengthwise direction of the slits occurs. By the measured difference in time of the signals at the first and second slit the deflection speed is determined.

Figure 7.235 shows the two different types of sensors. The left side shows a double-slit hole sensor, the right side a rotating sensor.

Another possibility for beam diagnostics is the application of the rotating wire sensor. The reason for the development of this new measuring variation in addition to the two methods described above lies in the question of whether and to what degree metal ions influence the power density distribution. In the case of a measurement with this sensor principle, a rotating tungsten wire with a diameter from 0.1 5 mm and a speed of up to 1000 s^{-1} moves through the beam. The deflection of the beam in this case is not necessary. The tungsten wire is coupled to a solid copper plate in order to increase heat dissipation. The current that is derived from the wire is measured in the form of a voltage signal. The measuring principle is comparable to the slit-measurement principle under the prerequisite that the diameter of the tungsten wire is smaller than the diameter of the beam.

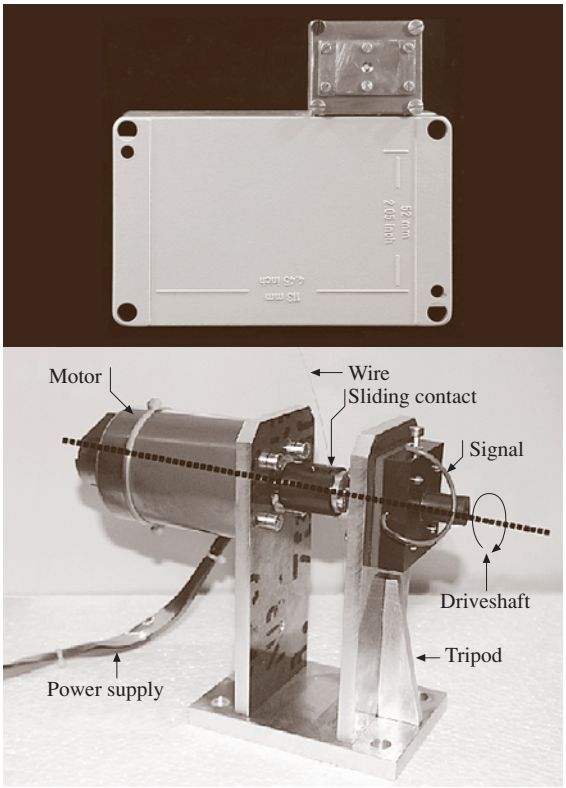


Fig. 7.235 Double-slit hole sensor and rotating sensor

Applications

Resulting from the large range of materials that are weldable with the EB (such as, e.g., tungsten, titanium, tantalum, copper, high-temperature steels, aluminum, gold) and material thicknesses, this method has a broad field of application possibilities, ranging from the specified microwelding of sheets with thicknesses of less than 1/10 mm, where low and extremely precise heat input is important, and thick-plate applications.

In heavy-plate welding, the process-specific advantages of the aforementioned deep-penetration effect and thus the joining of large cross-sections in one working step with high welding speed, low heat input, and small weld width begin to show. With modern welding equipment, wall thicknesses of more than 300 mm (aluminum alloys) and of more than 150 mm (low- and high-alloy steel materials with length-to-width ratios of approx. 50 : 1) are joined fast and precisely in one pass and without filler metal. In what follows, some application examples from different industrial sectors are given.

Below are some examples of industrial sectors with appropriate applications where the EB has become an established tool for material processing:

- Reactor construction and chemical apparatus engineering: welding of high-alloy materials, welding of materials with high affinity for oxygen, production of fuel elements and of circumferential welds of thick-walled pressure vessels and pipes
- Pipeline industry
- Turbine manufacturing: production of guide blades and distributors
- Aircraft construction: welding of structural/load-bearing parts made of titanium and aluminum alloys and of landing gears made of high-strength steels
- Automobile industry: welding of driving gears, pistons, valves, axle frames, and steering columns
- Electronics industry
- Tool manufacturing, e.g., manufacture of bimetal saw bands
- Surface treatment
- Material remelting
- Electron beam drilling with up to 3000 drills/s [7.206]

The numerous specific advantages of EB welding justify its increased application in industrial practice.

These advantagges include, for example, minimum workpiece heating by high power density, small beam diameters, and high welding speeds. The low heat input also allows for the welding of readily machined parts. The economic profitability of EB welding is not only a result of the high welding speeds and thus the short cycle times but also the high quality. The clean vacuum process without the influence of oxygen and the constant process parameters make the weld seams easy to reproduce.

These advantages are opposed by a series of process-related disadvantages. As the workpieces that are to be welded must be electrically conductive and as there is the risk of hardening and cracking due to the high cooling rates, the weldable range of materials is restricted. In addition, because the beam deflection is carried out via magnetic fields and the whole process needs to be shielded because of the development of X-ray radiation during welding, the investment costs are high.

7.4.4 Hybrid Welding

Emerging from the need to bridge larger gaps than was possible with the sole application of the laser and of welding faster than with a laser with filler wire addition, the hybrid welding methods have – also supported by major advances in laser technology in recent decades – been developed from its niche position to a method with a broad application spectrum.

Already in the 1970s, laser beam welding applications were widely used for their excellent process properties, which are due to the very high welding speed and low energy input. Due to production-related limitations, as, for example, rough part tolerances and large gaps, laser beam welding had not been used in many industrial sectors.

As a result of these problems and also because of the need to take advantage of laser beam welding, the idea of combining the laser with other welding methods arose by the end of the 1970s [7.207].

If a laser beam and arc are used in one production step for welding purposes, a distinction must be made between the combination and the coupling of the processes (Fig. 7.236). In the case of a combination, both joining tools act in different process zones and do not interact, except indirectly, such as preheating or tempering and similar effects. The two processes are combined when a given plate or root face thickness is welded with a tack weld or as a root pass by means of a laser beam and a subsequent SA or GMA welding [7.208].

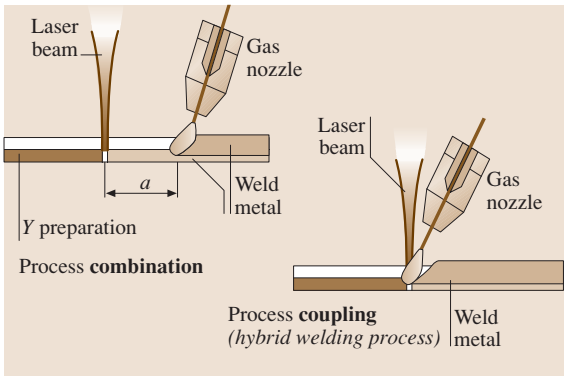


Fig. 7.236 Process variations of the combination of laser beam and arc

If both energy carriers are coupled in one common process zone, this is called a *hybrid process* or hybrid welding process. The coupling of both methods results in process-specific advantages called synergistic effects.

The laser beam and the arc interact in the process zone. They are coupled via the molten metal and in the majority of cases via the common process plasma [7.209].

In hybrid welding, the arc process effect of lowering the molten pool surface may result in an increase in weld penetration depth through the adaptation of the focus position [7.210, 211]. The weld penetration depth is mainly determined by the laser beam's power, power distribution, and shaping. The weld width is primarily determined by the arc, particularly the arc voltage. At a constant laser beam power, under special conditions an increase in the welding speed was achieved, a speed that was twice as high as the speeds obtained in pure laser beam welding [7.208, 212]. In contrast to the laser beam welding process, the energy efficiency of the whole process is increased through the application of an arc power source that works more efficiently than the laser.

Types of Hybrid Welding

The designation of the hybrid welding processes depends on the different process combinations and is stipulated in a guideline of the Deutscher Verband für Schweißen und verwandte Verfahren e.V. (DVS) (Table 7.34 [7.209]).

Laser Beam-GMA-Hybrid Welding Processes. In laser beam-GMA-hybrid welding, the addition of filler material is carried out via the arc process. The applica-

tion of GMA technology has the advantage that the filler wire can be molten with relatively inexpensive energy while the high-quality laser beam energy is maintained in an almost undiminished form for the transport of material into the depth. The positioning of the wire with respect to the laser beam is also relatively unproblematic in comparison with cold filler wire as the arc is positive effected through the common plasma and is pulled towards the keyhole (Fig. 7.237).

The arc is either pulsed via commercial GMA welding power sources [7.213] or triggered by direct current arc, in most cases a spray arc. The shielding of the welding point is, as in CO_2 laser beam welding, realized through pure helium [7.210] or helium/argon gas mixtures. Already low argon content in the shielding gas exerts a positive influence on the material transition and arc stability [7.212]. The applicable gas mixture is a compromise between high plasma shielding and arc stability. The admixture of low oxygen quantities for better drop formation and detachment is possible and also effective [7.214]. Welding filler materials are mainly of the same material as the base material and commercial solid wires that are also used for GMA welding of the base materials [7.215].

For thick-plate welding primarily CO_2 laser beam sources are used for their higher available powers. For welding of plate thicknesses > 10 mm, nowadays CO_2 lasers are applied as this thickness corresponds with the maximum penetration depth of modern Nd:YAG solid-state laser beam sources of up to 6 kW [7.216]. It is expected that, due to new developments, solid-state lasers will catch up to the CO_2 lasers wit respect to the available laser power in the next few years and thereby match their weld penetration depth.

Besides the square butt preparation, also V-type joint shapes and single V-butt joints with broad root face (Y) can be used; these are partly the result of cutting without any additional edge preparation. The energy of the laser beam is, in contrast with laser beam welding with cold filler wire, not needed for melting the filler material; this is – on the contrary – fed to the process already in a molten state, thus a reduction of the welding speed is not necessary [7.217]. The filler material allows for a defined, metallurgical influence on the welded structure [7.218, 219].

While in laser beam welding primarily parallel welds with a high aspect ratio are produced (Fig. 7.238a), a GMA weld leads to a low aspect ratio and reinforcement (Fig. 7.238b). The merging of the laser and the GMA weld leads to typical hybrid weld seam geometry. In particular the upper part of hybrid welds is widened (Fig. 7.238c), resulting in a triangular or *mushroom* and/or *bell-shaped* weld shape [7.217].

The GMA torch can be arranged in different ways relative to the laser beam. It is partly in the welding direction in front of the laser (backward-pointing GMA torch) [7.210, 220] or behind the laser beam (forward-pointing GMA torch) [7.221–224]. The *forward-pointing* arrangement is mainly used in hybrid welding of aluminum alloys in order to create an arc attachment point where the oxide layer has been removed by the laser beam, which again results in the significant increase in the process stability [7.225].

Laser Beam–Plasma–Hybrid Welding Processes. Through a suitable combination of the two heat sources of an electric arc and laser beam, a controlled influence is exerted on the heat input in the weld region.

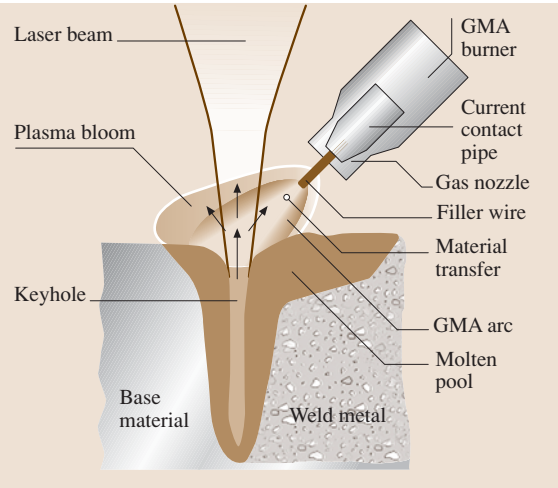


Fig. 7.237 Laser beam-GMA-hybrid welding process

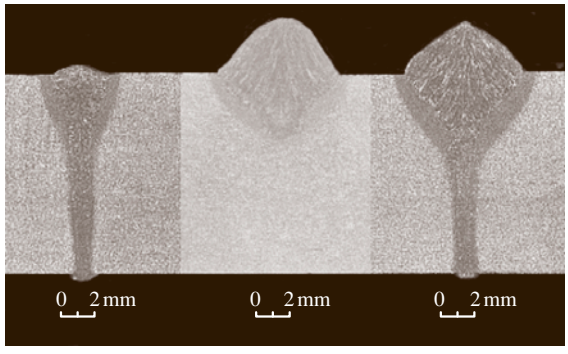


Fig. 7.238a–c Weld shapes: **(a)** laser, **(b)** GMA, **(c)** laser-hybrid

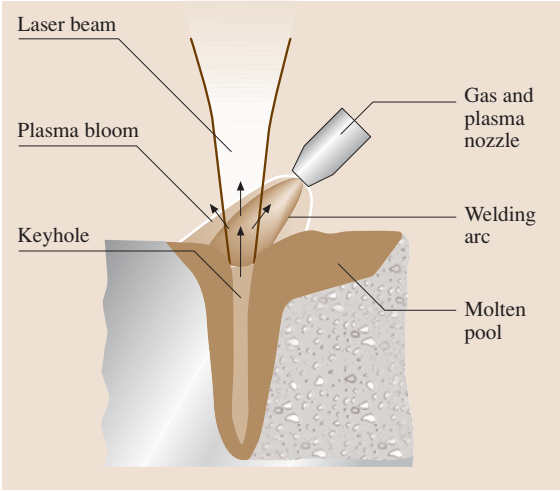


Fig. 7.239 Laser beam-plasma-hybrid welding process

Figure 7.239 shows the basic principle of such a process with a forward-pointing plasma torch. In contrast to pure laser beam welding, a plasma torch is arranged sideways or coaxially to the laser beam. When there is a need for filler wire, a cold wire can be added. Due to the plasma arc, the positioning of the wire is in comparison with a laser process with cold wire very easy.

The plasma arc brings about an additional heat input that can be independently adjusted by the laser beam. The electric arc essentially has three functions in this respect:

- It supports the beam welding process through an additional energy input and thus increases the operative capability and, as appropriate, the efficiency of the overall process.
- The concentric arrangement of an annular plasma arc around a laser beam brings about a defined heat treatment, in which respect the arc acts during the forward and backward run. The heat treatment provides the ability to reduce the cooling rate and thereby decreases the susceptibility to hardening and the development of residual stress states. It is, consequently, possible to adapt the microstructural condition to the application at hand.
- It produces a plane weld surface with a smooth and continuous weld interface.

Enhancing the operative capability of the beam welding processes through combination with plasma, arc technology opens up other new fields of application for welding, thereby presenting economic and manufacturing-related advantages that lead to savings in terms of fabrication time and costs and thus improve competitiveness [7.226].

The design-related benefits of this coaxial process variant – on the one hand, the small space requirement and, on the other hand, the nondirectionality of the process – are offset by the use of welding filler metal.

Laser Beam-TIG-Hybrid Welding Process. In laser beam-TIG-hybrid welding, the laser beam is, together with the TIG process, linked into the produced vapor cavity Fig. 7.240. One positive aspect of the use of a CO_2 laser is the developing plasma, which is generated by the laser process. The plasma is used as an ignition aid for the arc process and its effect is a considerable stabilization of the arc process in the hybrid welding process, especially in welding with low powers. [7.209].

Laser beam-TIG-hybrid welding is mainly applied in thin-sheet welding. The arc affinity for melting primarily the protruding edges makes the method most suitable for the joining of tailored blanks and for fillet welds on lap joints [7.209].

Applications

In 2000, laser beam-GMA-hybrid welding was the first hybrid method ever applied in industrial scale manufacture. Since that time, the Meyer Werft company, in Papenburg, Germany, has been using its hybrid unit with a 12 kW CO_2 laser (Fig. 7.241) for ship panel production.

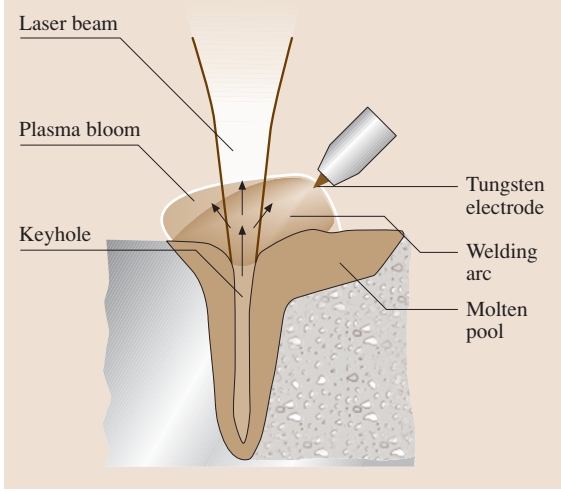


Fig. 7.240 Laser beam-TIG-hybrid welding process

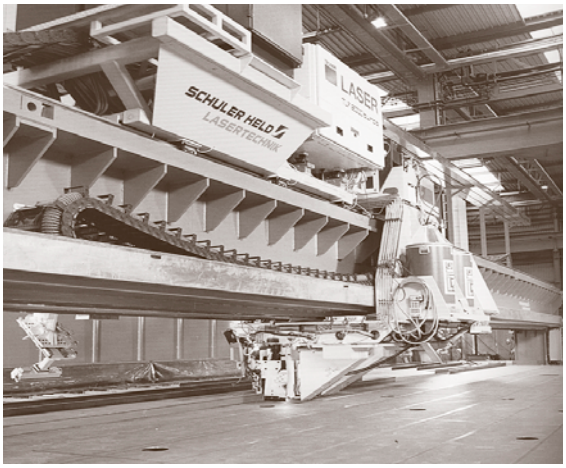


Fig. 7.241 Panel production with the hybrid welding method in the Meyer-Werft

The main advantages of the hybrid technology which led to its introduction in shipbuilding [7.227] are as follows.

- *High process speed* In contrast to conventional methods (SA for butt welds and MAG for fillet welds), it is possible to triple the welding speed. Also, even thicker panels can be welded in one part without reversing. In contrast with laser beam welding with filler wire, the welding speed can be duplicated. An additional advantage is the reduced throughput time and resultant less fixed material in the panel production.
- *Low thermal distortion* Due to the reduced energy input per unit length, the transverse shrinkage is, compared with conventional methods, reduced to a third. This leads to a significantly reduced amount of rework.
- *Good gap-bridging ability* Depending on edge preparation and plate thickness, air gaps between 0 and 1.5 mm can be welded. In contrast with the substantially lower values in laser beam welding, these tolerances can be realized with a justifiable expenditure, even with large parts in shipbuilding.
- *Excellent weld properties* Problems, such as, for example, high hardness and hot crack formation, as occur occasionally in laser beam welding have not been observed in hybrid welding, not even with large weld lengths.
- *Good process stability* Destructive and nondestructive tests showed that the correct balancing of the processes mainly results in consistently good weld quality. The results of the fatigue tests are shown in Fig. 7.242.

Laser hybrid welding has also become established in the automobile industry. The aluminum door frame of the Volkswagen Phaeton, for example, is manufactured by means of the hybrid weld process. This technique is also used for the Audi A8 space frame.

A further example of the efficient use of this hybrid technology is the joining of dissimilar material thicknesses, e.g., in the fabrication of tailored blanks. Figure 7.243 depicts the transverse microsection of a tailored blank. An example of the application of hybrid welded tailored blanks is the automotive floor panel with a thicker sheet in the tunnel zone. The low deposition volume of laser beam welds often leads to metal transfers with small radii. Apart from the notch-

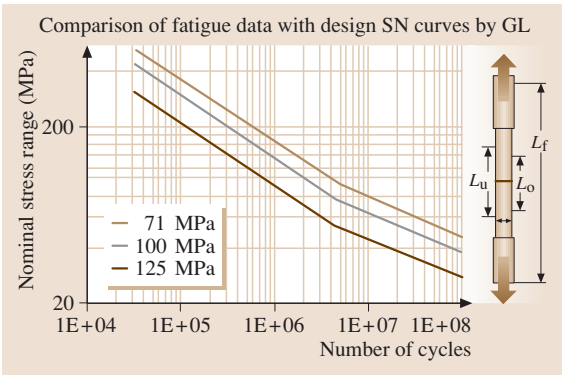


Fig. 7.242 Fatigue strength of hybrid welded butt joints (after [7.215])

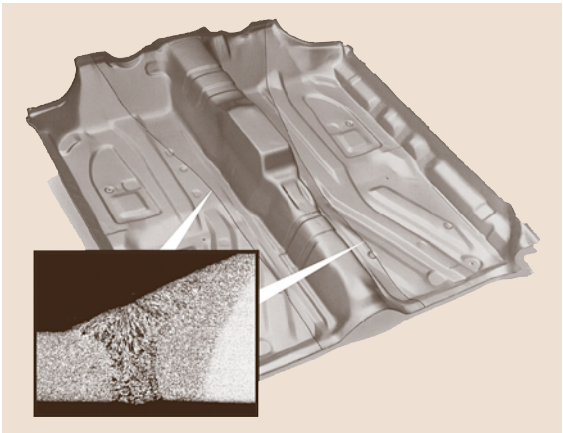


Fig. 7.243 Hybrid-welded blanks

Table 7.49 Process combinations hybrid welding (after [7.209])

	CO_2 laser	Nd:YAG laser
GMA	CO_2-laser beam-GMA-hybrid welding process	Nd:YAG-laser beam-GMA-hybrid welding process
Plasma	CO_2-laser beam-plasma-hybrid welding process	Nd:YAG-laser beam-plasma-hybrid welding process
TIG	CO_2-laser beam-TIG-hybrid welding process	Nd:YAG-laser beam-TIG-hybrid welding process

effect problem, such edges prevent the adhesion of applied coatings or paint films [7.228]. Use of the hybrid technology resolves such problems thanks to low-notch transitions produced by additionally fused material. Higher process speeds can also improve the productivity.

Summary and Perspectives

The advantages of the laser beam arc hybrid welding processes can be summarized as follows:

- High process speed
- Low thermal distortion
- Good gap-bridging ability
- Excellent weld properties
- Good process stability

Despite all the mentioned advantages of the laser beam arc hybrid welding methods, the fact still remains that, as a result of using two coupled welding methods, the number of setting variables also increases. For a reliable and stable hybrid welding process, the exact setting of the parameters is mandatory.

Laser hybrid applications in manufacturing, e.g., panel production in Meyer Werft ships or the manufacture of space-frame elements in the automobile industry, show the high potential of this method. Through the development of innovative laser systems, for example, the fiber laser and the disc laser, which excel with their compact design and high efficiency, new possibilities for the laser beam arc hybrid welding methods present themselves. The increasing power and beam quality of the solid-state laser will give rise to new possibilities, especially for welding light metals.

The increasingly compact design of the laser beam sources allows, for example, for mobile application in pipe production or application in the aircraft and aerospace industries.

7.4.5 Joining by Forming

Nonthermal joining technologies are used in almost every area of industry. Besides adhesive bonding and joining by pressing-on or expansion fit (e.g., threaded fasteners, clamping, pressing, wedging, cramping) joining by forming technologies is also significant and fundamental for multimaterial design in the sheet-metal-processing industry (Fig. 7.244).

Joining by forming processing, or mechanical joining as it is called in several publications, is enjoying a renaissance in its applications and its further development. As early as 2500 years ago in Greece pieces of cast bronze were joined permanently by rivets. Since that time especially arms, jewellery, and items of practical use have been joined or decorated by full or hollow rivets or with hemming or crimping processes. By the start of the Industrial Age riveting and crimping were being used in the production of steam boilers, e.g., at the site of railway construction and later in mechanical engineering and especially riveting for bridge-building as well. Since the 1930s, due to the costs involved, riveting was replaced by welding, particularly in bridge building.

In contrast to this development mechanical joining is now increasingly being used due to its many advantages. Modern high strength materials that acquire their mechanical properties by special heat treatments cannot be welded as traditional steel any longer and new joining methods without thermal influencing must be used to prevent weakening and high residual stresses of joined parts. Furthermore, many advantages (Table 7.50) justify the use of mechanical joining technologies.

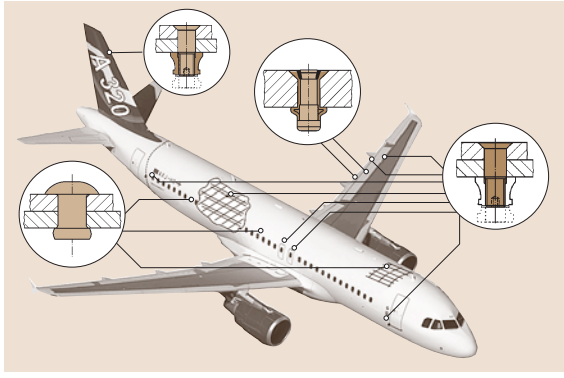


Fig. 7.244 Typical applications of joining by forming technologies at aircraft construction (after [7.229])

Table 7.50 Advantages and disadvantages of mechanical joining technology

Advantages	Disadvantages
● No thermal structural transformation of work pieces and therefore neither distortion nor residual stresses nor embrittlement ● Big variety of metallic and non-metallic materials and material combination can be joined as well as different material thicknesses ● Simple quality control mechanism ● High economic efficiency (small investment costs of machinery, rarely pre- and post-treatment of work pieces necessary) ● Materials with surface coatings can be joined without additional expenditure ● Very good environmental behaviour (neither emission nor pollution) ● High process reliability.	● Only lap joints possible, with redirect of power flow and higher weight of work pieces and therefore higher costs ● Lower tolerable static strain compared to e.g., welded joints ● Geometrical unevenness due to nature of processes (local protrusions of joining area) ● Usually more difficult correction and repair of improper joints ● Poor standardization and calculation methods

The broadest range of applications of joining by forming technologies is in the automotive industry just as well as in the field of heating, ventilation and air conditioning and the ship-building and container-building industries. Because of many new and specialized applications, additional or modified variants of mechanical joining technologies are emerging.

Overview

Terms. The technical term *mechanical joining* is generally used for technologies and processes for manufacturing permanent connections between two or more workpieces by transforming at least one of the workpieces or an additional fastener. Fasteners are additional parts needed for producing a connection with form or force closure between two or more workpieces, such as blind rivets. Some connections can only be produced by the use of fasteners, e.g., rivet connections; others do not need any additional parts, e.g., clinch connections.

The two ends of fasteners can be distinguished into a set-head and a closing-head. The set head remains after the joining process on that side from where the fastener is set into the workpiece, while the closing head – which is the opposite end of the fastener – permanently closes the connection.

Joining by forming of tubes, sheet metals, or sectional metals requires overlapping areas of workpieces. Different layouts of possible lap joints are shown in Fig. 7.245. Afterwards, all of these connection types will be called a lap joint. Butt joints cannot be joined by mechanical joining technologies.

Classification. So far no international standard for classification of joining by forming technologies exists. In German DIN standard several mechanical joining technologies are standardized at DIN 8593-5 [7.230] are classified by type and size of joined workpieces. Technologies are divided into joining of wires and tubes, sheet metals, or sectional metals. The joining technologies riveting and hydroform joining, which are classified by themselves, have an exeptional position, although they do not specifically differ from all other mechanical joining technologies of this standard.

An overview and classification of joining by forming technologies of sheet and sectional metals are shown in Fig. 7.246. Technologies shown in this classification are divided into joining with or without fasteners and also divided into during-the-process transformed parts, workpieces, and/or fasteners.

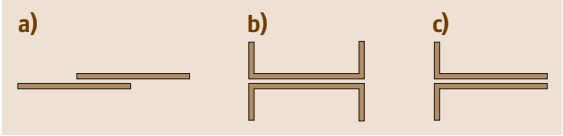


Fig. 7.245a–c Type of possible joints of mechanical joining technologies **(a)** lap joint; **(b)** parallel joint of U-sections; **(c)** parallel joint of flange connection

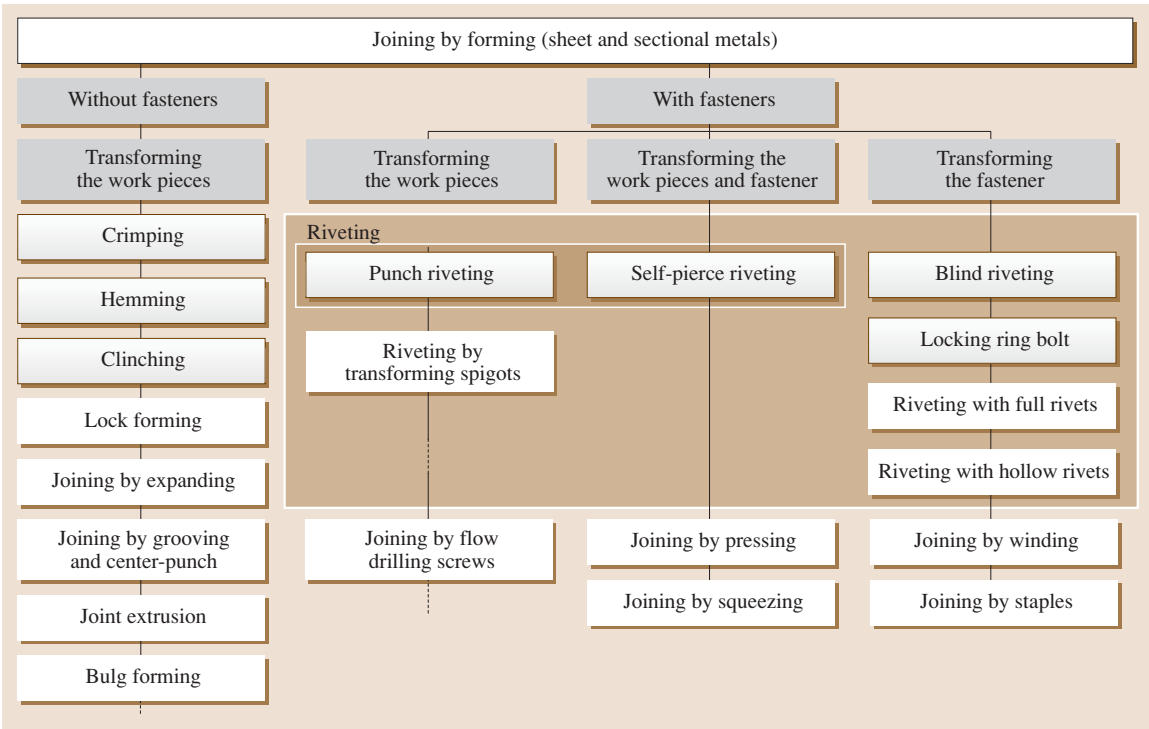


Fig. 7.246 Classification of joining by forming technologies (after [7.231])

Great significance for both applications and development is given to joining by forming technologies hemming and crimping, clinching, and riveting. For this reason these trends of mechanical joining technologies will be described in detail later on.

Selection of Significant Mechanical Joining Technologies

Hemming and Crimping. With hemming and crimping, overlapped edges of workpieces will be conjointly transformed to achieve a permanent form and force closure between the parts to be joined. Crimping used to be only a transforming process, more precisely a tension-pressure forming process, to turn up edges of curved sheet metals. Partial regions of convex or concave edges of workpieces are stretched or upset. For joining by crimping, the edges of containers, barrels, or tube ends will be transformed gradually to achieve a form and force closure between two or more parts (Fig. 7.247). In contrast to crimping, hemming describes a process where sheet metal edges are transformed in the same way to produce form- and force-closed connections (Fig. 7.248).

Various types of crimping and hemming technologies exist that require different kinds of tools and operation steps. Selection of a certain type has to be

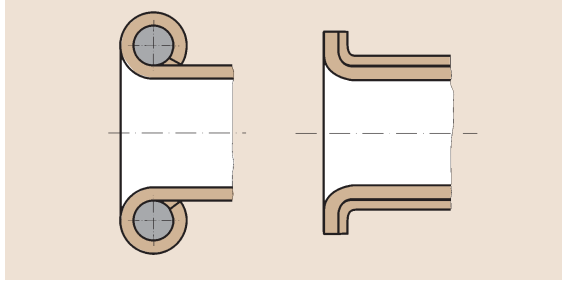


Fig. 7.247 Crimping joints

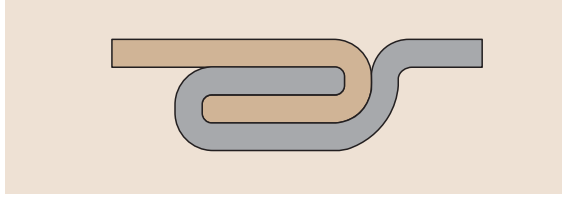


Fig. 7.248 Hemming joint

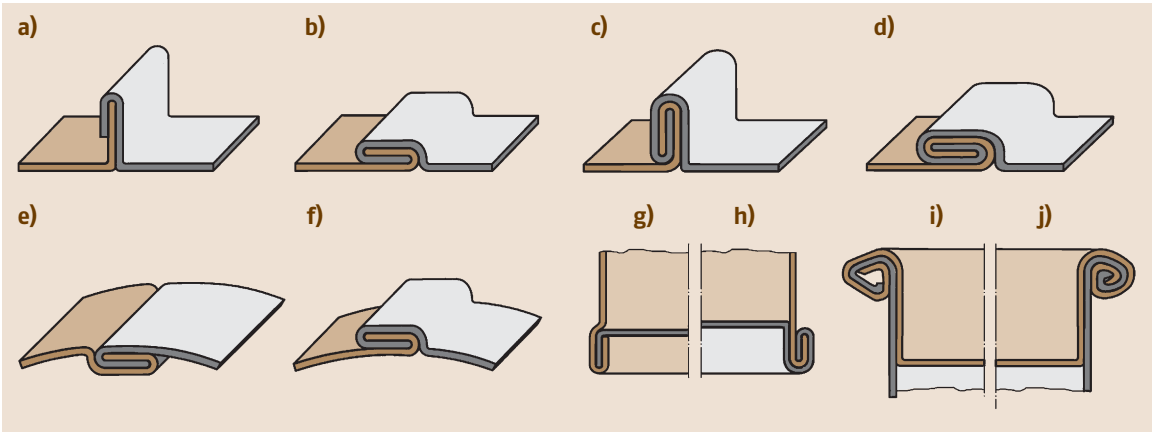


Fig. 7.249a–j Examples of different hemming and crimping joints (after [7.232]): **(a)** Stand-up hemming joint, **(b)** lying hemming joint, **(c)** stand-up double hemming joint, **(d)** lying double hemming joint, **(e)** inside hemming joint, **(f)** outside hemming joint, **(g)** single bottom crimping joint, **(h)** double bottom crimping joint, **(i)** trapezium crimping joint, **(j)** taper crimping joint

based on the kind of strain, accessibility of joining area, and other boundary conditions. Examples of different hemming joints are shown in Fig. 7.249.

Since there are no necessary surface preparations for hemming and crimping, parts only need to be aligned and fixed before. The manufacturing process can be divided into three steps:

1. Adjusting
2. Hemming or crimping
3. Redressing

With the first operation adjusting (1) the edge of all overlapping workpieces is bent up to a necessary minimum, though these hemming tools can grab the edge sufficiently. In this step, the size of the flange and hemming joint will be already fixed. The actual hemming or crimping (2) encompasses one or more gradual bending steps (Fig. 7.250). After closing the hemming or crimping seam, another redressing step (3) can plastically upset the joint and raises tightness and stress.

Hemming and crimping require a sufficient plastic formability of at least one workpiece to be joined, which can be joined with another nondeformable workpiece by form or force closure. In particular, materials with good cold formability are suitable for this joining method. One special field of application is the joining of different materials. Typical materials and material combinations are shown in Table 7.51.

Sheet thicknesses of workpieces joined by hemming are lower than 2 mm because of high forming degrees. Material thicknesses of tubes and barrels with crimping joints should be 6 mm at most depending on the level of mechanization. Te mechanical properties of hemming or crimping joints are determined by base material properties. Te connections of different materials are exposed to corrosion.

Clinching. Clinching is joining of lapped workpieces without any fasteners. A local limited area of material is pressed together by a punch into a die. A protrusion is formed by the materials yielding and creating a mechanical interlocking. The workpiece on the side where the punch presses into the material is called the punch-side piece and the opposite, which forms the protrusion, is called the die-side piece. Joining of more than two parts in one step is possible. In addition to form and force closure, under certain conditions a metallic continuity

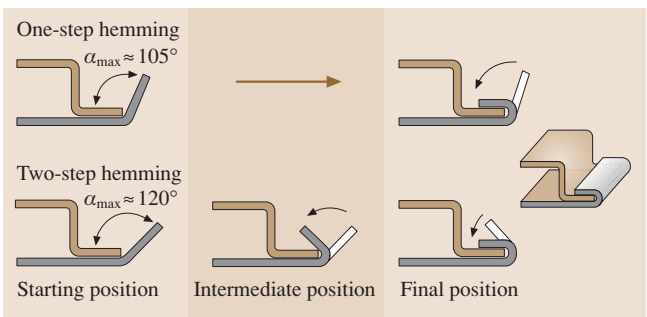


Fig. 7.250 Manufacturing steps of a hemming joint (one-step and two-step process)

Table 7.51 Materials and applications for hemming and crimping (after [7.233])

Materials and combinations	Applications
Copper	Metal fittings at roofs
Aluminum (e.g., EN AW-AlMg3, EN AW-AlMg4.5Mn)	Air conditioning and ventilation systems Sun roof and car doors
Steel (e.g., DC01, DC04)	Automotive applications
Bronze (also with other metals like silver, gold etc.)	Jewellery and art
Aluminum combined with steel	Automotive applications

(Fig. 7.251) can be found, which raises the connection strength.

Manufacturing of this kind of connection without any fastener is the major reason for its high economical efficiency, because of the saved costs of additional parts and the automatic feeding system of fasteners.

Clinching requires a double-sided accessibility. There is no need to remove coatings, like foils or paints, but for additional creation of metallic continuity clean surfaces free of dirt, oil, oxides, and coatings are needed. This could be reached, e.g., by cleaning with acid or especially for aluminum by deoxidizing or caustic etching.

The clinch tool consists of a punch and a die. The die can be only one massive part with a cavity and a ring groove (Fig. 7.252b) or have moving die blades (Fig. 7.252a) that are clamped together by springs. These die blades enable expansion of the diameter and allow an outward flow of metal to create a form- and force-closed permanent interlocking. A retainer fixes the workpieces before the punch is pressed into them and prevents distortion of the parts while joining is in progress.

Punch-and-die geometry influences the shaping of the interior and exterior geometry of clinch connections. Clinch technology can be divided into cutting and noncutting processes. Noncutting clinch processes produce no deliminations inside the connection. Material will be only transformed to reach form and force closure. For cutting clinch processes, punches with cutting edges are used that partially cut material and enable punch-side material to flow behind die-side material to produce an interlocking. A typical kind of cutting clinch connection is the square clinch because of its rectangular shape of protrusion. Besides dividing clinch systems according to their cutting ratio, they could also be classified according to their joining kinematics. Single-step and multistep clinch processes are known. Owing to many advantages (Fig. 7.253) mostly noncutting round clinch

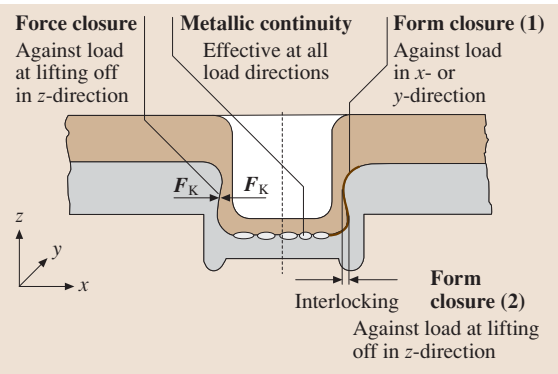


Fig. 7.251 Characteristic of clinch connections, e.g. round joint (non-cutting, single step process) (after [7.234])

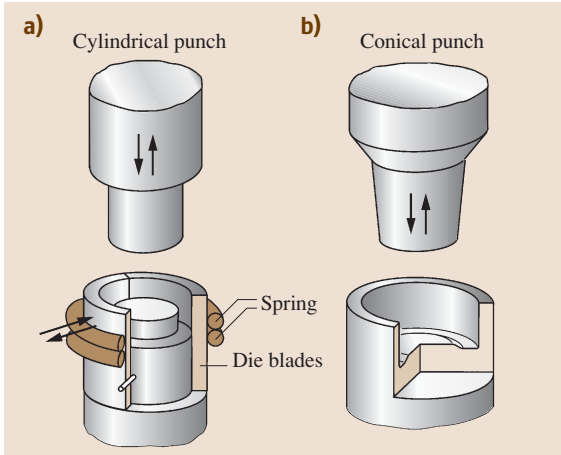


Fig. 7.252a,b Examples of clinch tool sets (after [7.235])

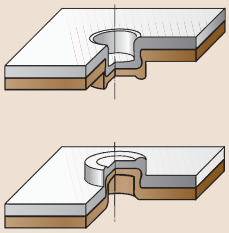
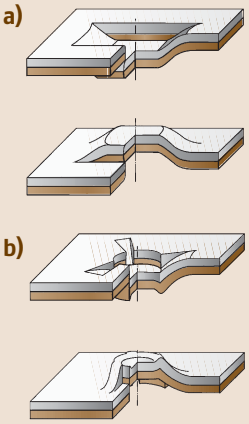
Non-cutting clinch connections	Cutting clinch connections
 Round joint	 Square joint (a) and special shape (b)
Advantages • Omni-directional load possible • Symmetrical look of connection	Advantages • Lower joining forces required compared to non-cutting clinch systems • Higher torsional stiffness of single connections
Disadvantages • Higher joining forces required • High strain hardening of material inside of connection	Disadvantages • Connection is not water-tight or airtight • Crevice corrosion is possible • Asymmetric load capability • No symmetric look of connection

Fig. 7.253 Comparison of non-cutting and cutting clinch systems

processes manufactured in a single action are used in European industry.

The principal working steps to produce clinch connections are almost identical for different kinds of clinch systems (Fig. 7.254). First, the parts to be joined are positioned between the punch and the die. Second, a retainer fixes the parts before the punch is pressed into

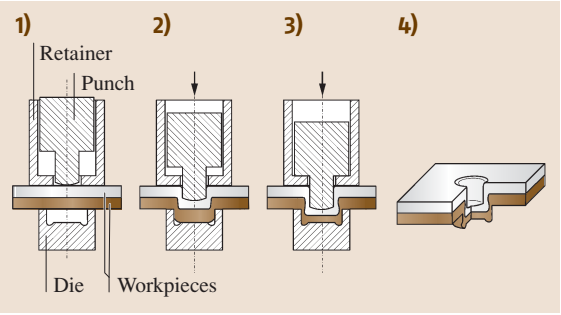


Fig. 7.254 Process steps of clinching (non-cutting, round joint)

them and creates a form and force closure by partial transformation. Using single-step processes, the punch will be drawn back after the preset bottom thickness of the clinch connection is achieved. For multistep processes one or more additional steps are required to produce an interlocking inside the connection.

Major applications of clinching are in the automotive industry, air conditioning and ventilation systems, as well as in the household appliance industry, e.g., washing machines, dishwashers, and refrigerators. This joining technology is used especially for lightweight constructions involving high-tensile steel and aluminum.

With clinch connections similar materials and material combinations can be joined. The suitability of materials depends mainly on deformability. The size of deformability can be measured by elongation on break, the yield ratio, and the tensile strength of the base materials. Conditions for good and restricted suitability for clinching are shown in Table 7.52. Restricted materials for clinching have to be proofed in particular cases.

Well-suited materials for clinching are naturally hardened and age-hardenable aluminum alloys, steel alloys for cold working, mild steel, cold-rolled dual-phase steel, high-tensile steel, ferritic and austenitic stainless steel, as well as copper and bronze alloys. Materials with an overall sheet thickness of up to 6 mm can be joined by clinching.

Table 7.52 Mechanical properties of well-suitable and restricted suitable materials for clinching (after [7.233])

	Well-suitable	Restricted suitable[a]
Elongation at break	$A_{80} \geq 12\%$	$12\% > A_{80} \geq 8\%$
Yield ratio	$\frac{R_{p0.2}}{R_m} \leq 0.7$	$\frac{R_{p0.2}}{R_m} > 0.7$
Tensile strength	$R_m \leq 500\,\text{MPa}$	$500\,\text{MPa} < R_m \leq 700\,\text{MPa}$

[a] Have to be proven separately

Because of form and force closure clinch connections should be preferentially loaded with shear stress. Head and peel tests of clinch connections show much lower possible loads. Torsional stress should be generally avoided for single spot connections, such as clinch connections.

Thermal, chemical, and electrical properties of clinch connections depend on properties of the base materials used. A combination of different materials raises the risk of electrochemical corrosion.

Riveting

Blind Riveting. In the early 19th century new manufacturing technologies and fields of applications required the development of joining technologies with only one-sided accessibility but comparable mechanical properties as full rivets used so far. Joining by accessibility of only one side is called *blind*, and so this newly developed kind of fastener was also called *blind rivet*, which later became very significant for aircraft and vessel construction. Besides the economic advantages of faster manufacturing rates, blind riveting differs from riveting with full rivets by a high noise reduction and less damage to the workpiece surface.

The development of different types of blind rivets took place gradually and many different kinds were invented. After trepanning full rivets, which still needed two-sided accessibility, by the early 20th century, for the first time blind rivets consisting of two parts were being produced in England. In Germany, not before the 1930s, blind rivets were developed that used a small explosive charge within a cavity to create a closing head after ignition by a thermal influence on the set head. With this the closing head of the explosive rivet was expanded and closed the connection (Fig. 7.256a).

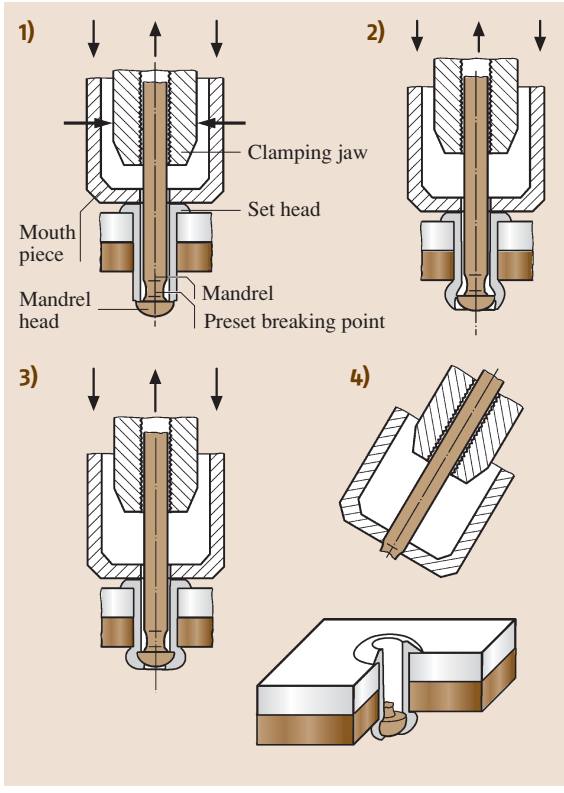


Fig. 7.255 Process steps of blind riveting (after [7.236])

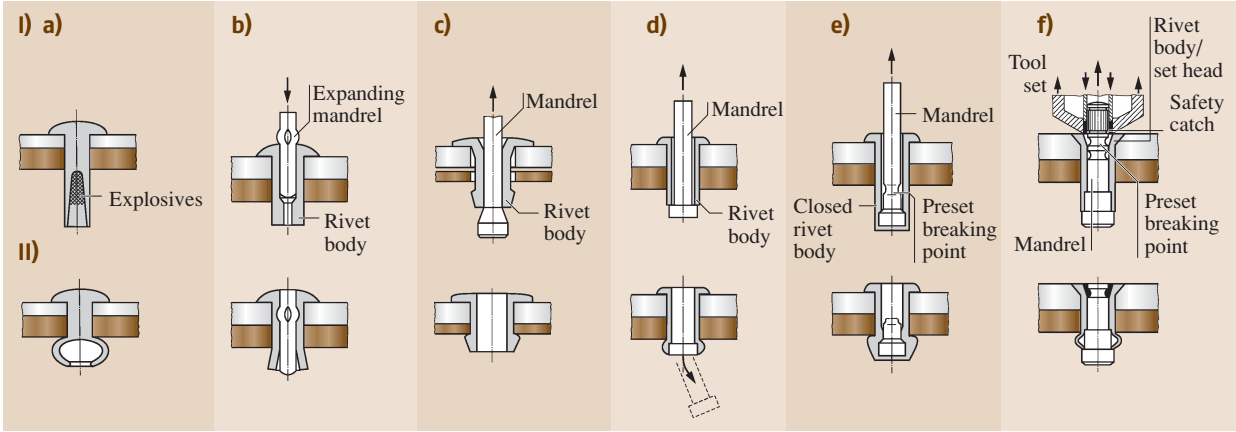


Fig. 7.256a–f Selection of different blind rivet types and schematic process of joining (**I** point of origin transforming closing head; **II** joined connection): **(a)** explosive rivet; **(b)** expanding rivet; **(c)** pull-through blind rivet; **(d)** blind rivet with loosen mandrel; **(e)** blind rivet with closed rivet body; **(f)** planar breaking countersunk-head blind rivet with mechanical mandrel safety catch

Along with other developments blind rivets were developed consisting of two or more parts that have different mechanical properties and requirements. Most common blind rivets have two parts, the rivet body and the mandrel, which form the closing head.

Blind riveting does not have special requirements for surface preparation like all mechanical joining processes. Different materials can be permanently joined with blind rivets. Releasing the connection is only possible by the destruction of the blind rivet. All parts need to be congruent, predrilled, and fixed to achieve joining with blind rivets. Holes can be drilled, punched, or brought in by another prior manufacturing step. The size and tolerances of the hole depend on the diameter and type of blind rivet.

Process steps are very similar for different types of blind rivets and are shown in Fig. 7.255. At first, the blind rivet is introduced into the setting hole and the rivet head is pressed against the set-head material. The mandrel is pulled by the clamping jaw of the setting tool and slips into the rivet body. The workpieces are pressed together by an increasing pulling strength. With this the mandrel head transforms the rivet body and creates a closing head. Depending on the kind of rivet, either the mandrel is pulled completely out of the rivet body or the mandrel breaks at the preset breaking point and is extracted by the setting tool. The required breakload will be determined by the cross section of the mandrel preset breaking point and by the material of the mandrel. Some kinds of rivets surround and fix the mandrel head and prevent it from loosening with vibrating loads. Different types of blind rivets differ in their required joining forces, connection strength, and mechanical properties as well as in costs associated with manufacturing the rivet connection.

Blind rivets were developed by the necessity of joining comparable connections of full rivets but having only one-sided accessibility. For that reason blind rivets can join almost every material and material combination without having special requirements of deformability or other material properties. The only precondition is that joining parts must have a premanufactured hole. Important for a good quality of blind rivet connection is the right choice of rivet diameter and rivet length for the joining task.

Since the mandrel has to transform the rivet body, it should have a higher tensile strength. For that reason, the mandrel and rivet body mostly consist of different materials. Although rivet bodies used to be made of aluminum because of its better deformability, nowadays also mild steel, stainless steel, and nickel and copper alloys are used to match the material of the workpieces and prevent corrosion. Blind rivets can be provided with different coatings.

Some criteria on the selection of materials for blind rivets are as follows:

- Avoidance of exposure to corrosion of workpieces
- Connection strength
- Cost

Most relevant applications of blind rivets are parts where a two-sided accessibility is not possible, such as closed sectional profiles, e.g., interior pressure-shaped profiles or large-area parts where connections should be set far away from the edges. But blind rivets mostly have a lower shear connection strength than punch rivets or locking ring bolt systems, especially those types of blind rivets where only the rivet body supplies the power between the workpieces.

The mechanical properties of blind rivet connections depend on the properties of base materials and the blind rivets used. Overall sheet thickness of workpieces to be joined can vary in a wide range from 0.5 mm to more than 20 mm.

Punch Riveting. Until the development of punch riveting for joining two or more parts it was necessary to have premanufactured holes for setting the full or hollow rivets. Holes could be set by drilling, punching, or some other thermal or nonthermal cutting process, such as watercutting. These additional working steps were problematical for increasingly automated production. In particular, hole-congruent positioning of more than two workpieces required high manufacturing precision. The additional process steps reduced the economic efficiency of the whole joining process.

With punch riveting workpieces are directly form and force closed and permanently joined without premanufacturing a hole. The cylindrical solid rivet produces the necessary hole itself while joining by punching a small slug out of all workpieces to be joined. The rivet itself is not deformed. Only parts are deformed locally, because the die-side material needs to be deformed in one single or multiple ring grooves of the punch rivet. Workpieces need to have two-sided accessibility for joining. Transmission of connection strength between joined parts happens exclusively by the punch rivet.

Because of its simple technology and less surface damage around the joining spot, especially on coated and painted parts, this connection is an economical alternative to resistance spot welding.

The tool set for punch riveting consists of a ring-shaped die, a retainer, and a punch (Fig. 7.257). The die is ring-shaped to enable throw-out of the slug while joining.

While the tool set moves to the joining spot, the punch rivet will be fed into the retainer and positioned to the workpiece surface. The retainer fixes parts before joining and guides the punch rivet until the joining starts. The punch presses the punch rivet into the work pieces. The punch rivet will be moved till the countersunk set head of the punch rivet is installed flush with the surface of the punch-side workpiece. The ring-shaped die has a small offset at its inner diameter, which causes the material of the die-side workpiece to flow into the ring groove of the punch rivet. With this the punch rivet is fixed permanently to the workpiece and the connection is established.

The joining process is distance controlled, meaning the stroke of the punch is defined by an adjustable stop collar. The required joining force is automatically adjusted and only limited by the nominal load of the riveting machine.

Materials of punch rivets and workpieces can differ, but the material of the rivets needs to have a higher tensile strength and hardness than the material of the workpieces, because the cutting lip of the rivet needs to punch holes into the parts. Similar materials or material combinations can be joined by punch riveting, but the die-side material must be deformable, so that the ring groove can be filled with base material.

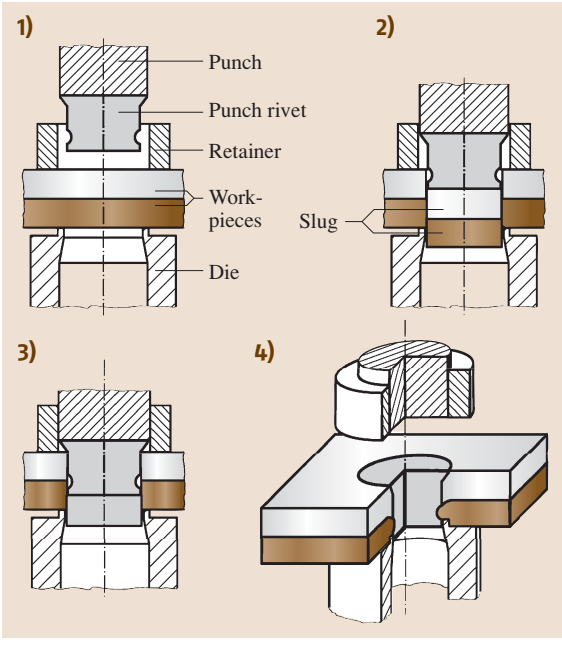


Fig. 7.257 Process steps of punch riveting (after [7.237])

Typical materials of the punch rivets are mild and stainless steels as well as aluminum, but most of the rivets are heat-treated steels. Since the rivet will not be deformed, higher hardness and lower deformability of rivet materials is possible; therefore, martensitic (stainless) chrome steel can be used too.

With punch riveting two or more parts can be joined at one time between an overall sheet thickness of 1 – 14 mm. These values are lower for high-strength steel.

Mechanical, thermal, chemical, and electrical properties depend on the base material and the punch rivets used. Punch-riveted joints commonly have a higher connection strength than, e.g., clinch connections because of their special force transmission with additional fasteners.

A material combination of punch rivets and workpieces can lead to electrochemical corrosion and should be prevented by coatings or paint of the joining area.

Self-Pierce Riveting. The reasons for and steps of development of self-pierce rivets took place in the same way as punch rivets. In contrast to a solid punch rivet, for self-pierce riveting partly trepanned cylindrical rivets are used. These self-pierce rivets expand while joining to a form and force closure. Both workpieces and self-pierce rivets are partially deformed at the joining process. The first self-pierce rivets punched workpieces completely, but current types of self-pierce rivets only punch the set-head side of workpieces and only deform the die side, where they create an interlocked close head. With this an air- and watertight connection is created without cutting and less damage at the die side. Self-pierce riveting requires a two-sided accessibility of workpieces.

Process steps differ just a little in comparison to punch riveting (Fig. 7.258). Parts to be joined will be positioned between the tool set, the punch, and the die and will be fixed by the retainer. Now, self-pierce rivets are automatically fed inside the retainer just under the punch. The connection will be set with one stroke of the punch. In contrast to punch riveting, here a die with a cavity is used that will be completely filled with material at the joining process. Self-pierce rivets also have a cutting lip, which punches a slug into the workpiece of the set-head side. This slug will be enclosed inside the rivet. The cylindrical shape of the die with a frustoconical punch in the middle of the

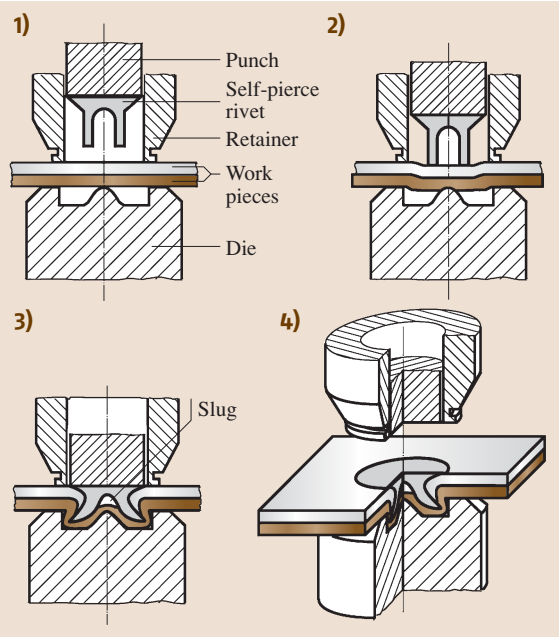


Fig. 7.258 Process steps of self-pierce riveting

cavity causes the self-pierce rivet to expand and creates a form- and force-close interlocking within the die-sided workpiece. For this reason it is necessary to adjust the length of the self-pierce rivet at the overall thickness of the workpieces and the die depth. In contrast to a punch rivet connection, the self-pierce rivet connection forms a protrusion at the die-sided workpiece.

In contrast to punch riveting, self-pierce riveting requires special properties of the rivet material. The rivet has to have a high enough tensile strength and hardness to punch into the set-head side of the workpieces and must be deformable by the frustoconical punch in the middle of the die without breaking or tearing.

With self-pierce rivets metals and nonmetal materials can be joined, but die-sided materials must have a sufficient malleability to be transformed into the cavity.

Workpieces can be joined with lap joints, where the minimum lap size depends on the size of the used self-pierce rivets and the tool set. More than two parts can be joined with an overall sheet thickness at a range of 1–6 mm [7.238]. High-strength material may reduce the maximum thickness that can be joined.

The properties of self-pierce riveted joints depend on the mechanical, thermal, chemical, and electrical properties of joined working pieces and the rivet materials used. To prevent electrochemical corrosion, parts and rivets should be coated or painted after the joining process.

Locking Ring Bolt Connections. Joining with locking ring bolts is a special kind of riveting. The locking ring bolt, also known as locking ring bolt system, consists of two components – a ring and a bolt. Locking ring bolt connections have unique properties. These connections possess high strength like full rivets and the possibility to join high-strength materials, similar to preloaded screws. Originally, locking ring bolt connections were developed for aircraft and aeronautics but are used extensively in the fields of structural-steel erection, cranes, rolling stock, utility vehicle manufacturing, and shipbuilding. In contrast to screws, the original preload of the connection is permanently retained. Therefore, locking ring bolt connections are vibration resistant. The parallel grooves of the bolts avoid preload reduction, even in case of vibrating load.

Locking ring bolt connections are permanent. The only way to separate the joined parts is to destroy the locking ring bolt. Prepunched parts and accessibility from both sides are prerequisites. Ordinary locking ring bolts have parallel grooves (Fig. 7.259). But there are variants with threadlike grooves and bolts with additional regular thread to join other components with nuts.

The joining process of ordinary locking ring bolts is shown in Fig. 7.260. The process can be divided into the following steps:

1. The bolts are plugged into the punched parts, and the ring is slipped over the bolt. The manufactured head and the ring are in contact with the parts. The rivet tool is fit to the pull zone of the bolt.
2. The clamping jaws of the rivet tool fix the bolt. The die of the rivet tool set is pressed against the ring and the workpieces.

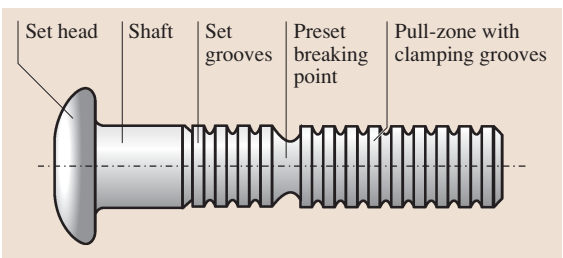


Fig. 7.259 Sections of a locking ring bolt

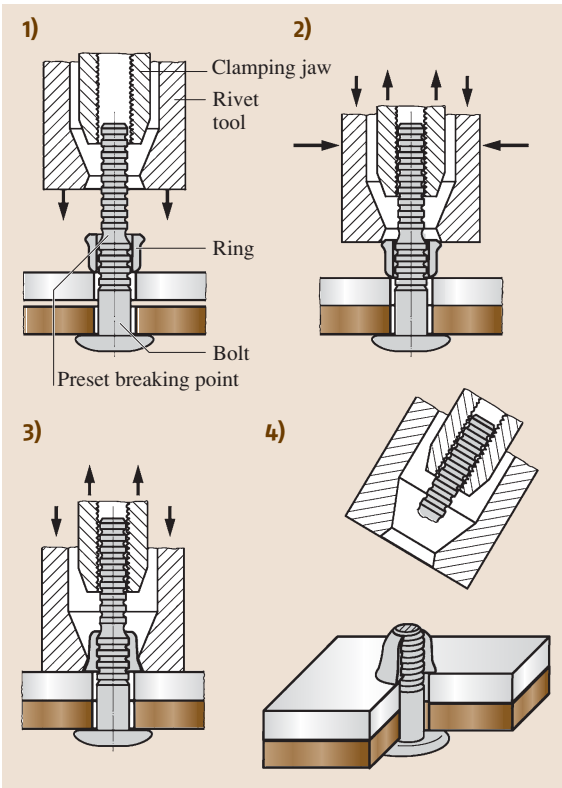


Fig. 7.260 Process steps of locking ring bolt technology (after [7.239])

3. The die of the rivet tool is moved continuously toward the manufactured head. In this way, the ring is formed into the grooves of the bolt. Axial pressure is balanced by clamping of the bolt. With this additional stress to the parts can be avoided.
4. The bolt is loaded by the tensile force of the clamping jaws until the bolt breaks at a preset breaking point. All tensile forces are balanced inside the tool. There are no loads to the workpieces.

All materials that allow manufacturing of holes by punching, drilling, and other processes are able to be joined by locking ring bolt connections. There are no special requirements for the materials regarding formability and other properties. The material of the locking ring bolt has to be adapted to the material of the parts in order to reach an optimum in strength and corrosion resistance. The fracture area of the bolt at the preset breaking point should be coated, unless the bolt is made of stainless steel or aluminum alloys.

As a result of the high-level strength, locking ring bolt connections are suitable for parts with overall thicknesses of more than 3 mm as well as materials of higher strength.

Pre- and Posttreatment

The outstanding features of mechanical joining are its robust and simple technology. Processes are very tolerant of varying material properties and especially workpiece surface preparation. Depending on the specific technology (e.g., for blind riveting) congruent pierces by hole punching, drilling, or water cutting can be required. Surface preparation is mostly unnecessary. Workpieces with coated and raw surfaces can be joined by most joining by forming technologies.

Posttreatment of mechanical joined workpieces is usually not required either. In case of surface coatings of joined parts it should be determined by inspection whether the surface was damaged by the joining process. Since process-required pierces are manufactured before joining, coatings can be restored before joining as well. In the case of used fasteners, corrosion prevention must be investigated.

General Topics of Quality Inspection

Quality inspection includes among other things inspection of the manufacturing process and the inspection of connections. The inspection procedures of the manufacturing process can be automated. For this, significant process parameters are measured using sensors. The actual indication and ideal indication are compared. Using this method, it is possible to assess the quality of a connection during the manufacturing process and to classify as required the quality level. As a consequence of increasing requirements on quality inspection, these inspection systems are common not only in automated manufacturing processes but also in partially automated ones. The most important parameters of joining by forming are the force of male die and its length of stroke, which can be monitored online. It is possible to draw curves of actual force vs. actual stroke and to compare these curves with ideal curves. The data of the actual curve are compared with the separate data of the ideal curve using tolerance ranges (Fig. 7.261a) or with overall data using tolerance bands of the parameter (Fig. 7.261b).

Connections can be proofed after manufacture by visual inspection, measuring of the bottom thickness (for clinch connections), or destructive tests [7.240]. Visual inspection is used to determine cracks on the

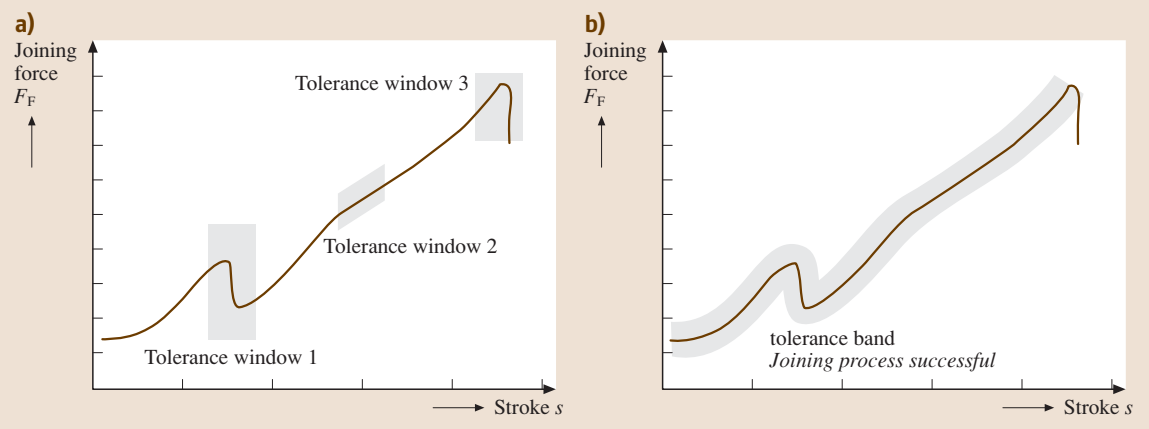


Fig. 7.261a,b On-line quality inspection of clinch connections with (a) tolerance ranges, (b) tolerance bands

surface, the symmetrical shape of the connection, and surface failures. The bottom thickness is an important property because other geometric parameters such as undercut and neck thickness relate to the bottom thickness for the used tool set. Until now, the strength of a connection was determined with tensile shear tests resp. peel test recording of the maximum forces reached.

Metallography enables one to investigate the inner structure of mechanical joined connections. For this, macrosections are useful. For dot-shaped connections, the section plane should divide the connection symmetrically. The dimensions that are responsible for the properties of the connection can be measured. Figure 7.262 shows the cross section of a clinch connection with characteristic dimensions.

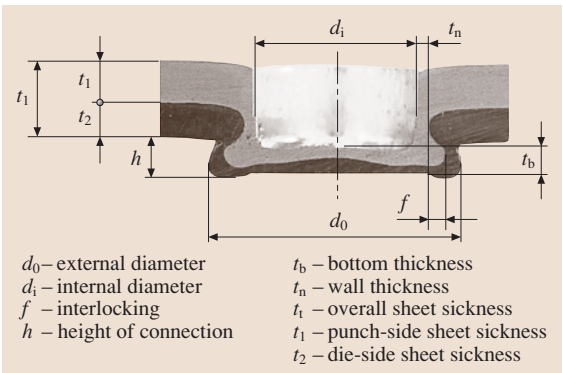


Fig. 7.262 Metallographic evaluation of clinch connection showing characteristic dimensions (after [7.233])

Safety in Operation and Environmental Protection

In the field of mechanical joining, the general instructions for machine operation and safety codes must be kept. Because presses or presslike machines are used for mechanical joining, the special safety codes for these machines have to be observed. In particular, high-speed moving machine components, e.g., in common use for setting full rivets, can cause dangerous noise. In this case, noise abatement or noise barriers are necessary.

Regarding environmental protection, all processes of mechanical joining can be estimated as being favorable. Emissions, such as fumes, gases, or hazardous substances, are not released. Most process steps do not need any surface treatment. Thus, no pickles or detergents are necessary.

7.4.6 Micro Joining Processes

Microstructure technology is used for the fabrication of microstructures and microcomponents, whose dimensions lie in the micrometer range and extend to the nanometer range. Microsystems technology is the consistent further development of microelectronics toward nonelectronic fields [7.241]. It links miniaturized, microstructured components and suitable sensors as well as actuator technology to form intelligent systems.

Microsystems technology makes use of techniques used in microelectronics, micromechanics, and microoptics as well as in mounting and connecting technology. Figure 7.263 shows the single procedural steps for the fabrication of a microsystem.

Microsystems have a broad range of applications, e.g., in microelectronics, in air and spacecraft, chemistry, biology, and medicine. The application of novel

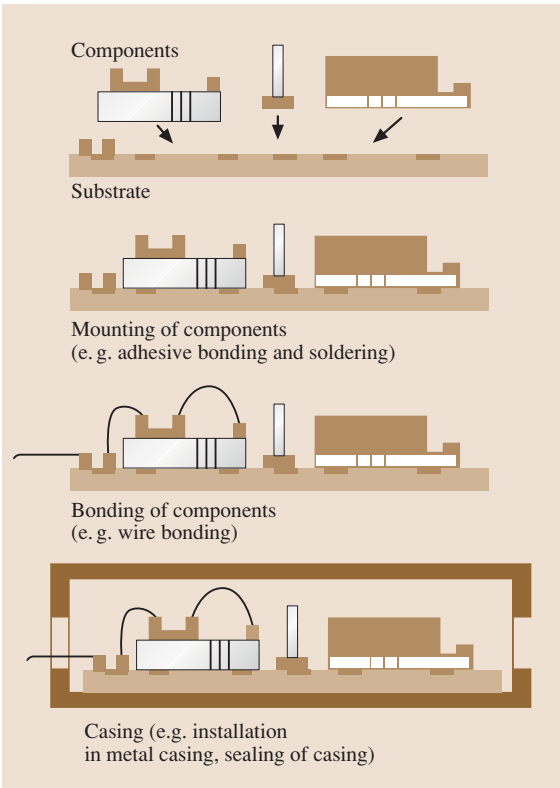


Fig. 7.263 Layout of a microsystem (after [7.243])

materials and the adaptation of manufacturing techniques in the field of miniaturization have decisively contributed to the success of microsystems technology. Within this framework, hybrid constructions have become of particular importance. The development of a hybrid, i. e., nonmonolithic construction in microsystems technology with single components from different technological fields, requires an adapted microjoining and processing technology [7.242].

Microwelding Processes

Among the microwelding techniques, pressure welding methods are of particular importance and most frequently used due to the advantages they offer. However, for special applications and the individual fabrication of precision parts fusion welding is used, too.

Resistance Welding. The majority of resistance welding techniques can be applied to microwelding processes. By means of transistor-controlled direct-current and medium-frequency inverters, welding times from 10 µs are programmable in 10 µs intervals by which an exact dosage of the welding energy required is achieved.

Spot Welding. Spot welding for the manufacture of extremely small and microwelded joints is applied for components used in the electronic, optical, and precision mechanics industry. This applies mainly to wires, foils, and strips made from different metallic materials and their alloys. To date, more than 170 combinations of the most important metals have been welded.

Seam Welding. Seam welding is used for the welding of cases of electronic elements and circuits. It is also applied for the manufacture of cushion-type membranes (bell-shaped and flat membranes), overlay-plated and inlay profiles for bimetal contact strips.

Projection Welding. Projection welding is used for the welding of wires and structural parts made from different metallic materials, and for the fabrication of hollow bodies for electronic components, sensor casings, contact bars, plugs, joint bars, and litz wires.

Parallel Gap Welding. Parallel gap welding is a process variant of resistance spot welding and comparable with indirect spot welding [7.244]. Figure 7.264 shows the typical arrangement of the electrodes. Both electrodes are lowered onto the piece to be welded from above at a very small distance.

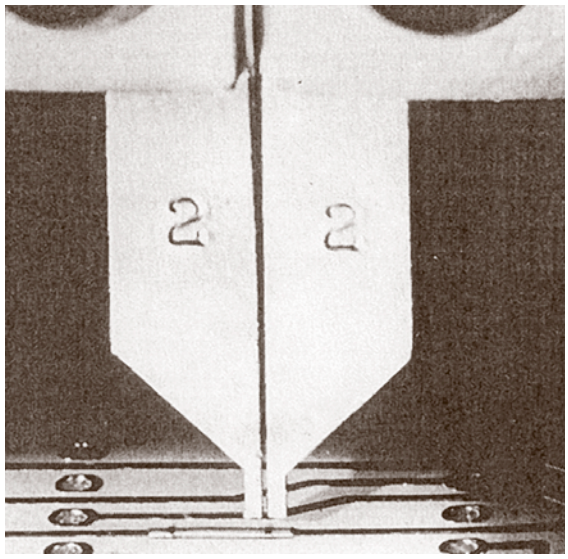


Fig. 7.264 Parallel gap welding used for repair work of strip conductors (after [7.244])

The electrodes are clamped in adjustable fixings. The gap width depends upon the material used and is continuously adjustable from 0.025 to 1.00 mm. Welding currents from 100 to 1000 A, welding times from 1 to 100 ms, and electrode forces from 0.2 to 100 N are used.

By using parallel gap welding, silver contacts of solar cells made from silicon (thickness 180–250 µm) are connected with silver strips (20–35 µm) to form batteries. 25 000–65 000 solar cells form a battery that is used for the direct transformation of sunlight into electrical energy to generate energy for spacecraft and satellites [7.245]. Further applications include the welding of thin wires and strips onto circuit elements, the bonding of sensors, and the repair of tracks of printed circuit elements with bonded strips from kovar, copper, silver, or gold.

Wire Bonding. Wire bonding is a highly developed microwelding technology and guarantees reliable joining in the micro range. Wire bonding is the most important joining technique for the manufacture of conductive welding joints in microelectronics. It is used for the manufacture of electric connections between semiconductor chips, structural components, and cases (metal, ceramics, and plastic casings). Bonding is accomplished by using a metallic wire loop that is welded at its ends onto the connecting metallizations of the components to be joined. In addition to lateral bridging, a difference in height can be bridged as well [7.246].

Welding techniques that are used for wire bonding differ with respect to:

- The manner of energy input
- The shape of the bonding tool
- The wire diameters applied

All techniques have one thing in common, i. e., joining is accomplished in the solid state of the materials (solid-state welding) [7.247].

Thermocompression Bonding. The joining of bond wire and connective metallizations is produced by the input of heat (150–350 °C) and power (0.1–0.9 N) using welding times from 0.3 to 0.6 s. Commonly, the substrate is preheated by means of a hot plate to about 150 or 170 °C. The bonding tool heats the area to be joined to the required process temperature in pulsed mode. Predominantly thin bond wires made from gold (wire diameter from 7 to 50 µm) are welded onto the bond pads of the chip and the wire bond face of the casing.

Ultrasonic Bonding. Bonding is produced by the effect of mechanical vibrations and the bonding force which are introduced by means of the bonding tool (operating frequency of 60 kHz, ultrasonic power from 0.1 to 30 W, vibration amplitude from 1 to 2 µm). Bond wires made of AlSi1 and AlMg1 (18–50 µm) are mainly used; for high-performance structural elements also wires made of highest-grade aluminum with a diameter of 100–500 µm.

Thermosonic Bonding. Thermosonic bonding is a combination of thermocompression bonding and ultrasonic bonding which unites the advantages of both techniques at low temperatures (100–180 °C) and shorter welding times as well as bonding forces. It is standard procedure used for the contacting of integrated circuits (ICs), single diodes, and transistors on ceramic substrates with thick-layer and thin-layer circuits. Gold wires with a diameter of 17 to 100 µm are preferably used.

For wire bonding, ball–wedge bonding (Fig. 7.265) and wedge–wedge bonding are applied (Fig. 7.268).

The manufacture of ball–wedge bonds can be accomplished by both thermocompression and thermosonic bonding. Due to the special shape of the tool

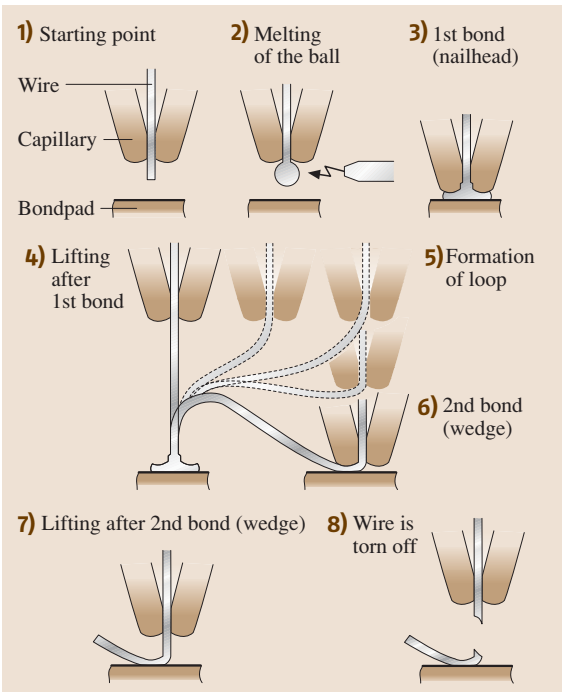


Fig. 7.265 Procedural scheme of ball–wedge bonding

Fig. 7.266 Nailhead-joint (after [7.248])

(bonding capillary), wire bonding can be accomplished in each direction without turning the tool or structural element. The meltdown of the ball is only possible by using gold wire (diameter 12.5–50 µm). Figure 7.266 shows the first bond (nailhead) and Fig. 7.267 the second bond (wedge).

In the case of wedge–wedge bonding (Fig. 7.268), the formation of the first and the second welding shows a similar geometry (Fig. 7.267).

Thermocompression, ultrasonic, and thermosonic bonding are suited for the manufacture of wedge–wedge joints. Bond wires and thin aluminum, copper, and gold strips can be used. Due to the shape of the bond wedge, the direction for the second bond is already defined by the formation of the first bond.

Automatic wire bonders can bond up to 18 wires/s using the ball–wedge technique and up to 5 wires using the wedge–wedge technique [7.251].

Laser Beam Welding. In the field of microjoining technology mainly pulsed Nd:YAG lasers operating with wavelengths of $\lambda = 1.064\,\mu m$ are used. The pulsing of

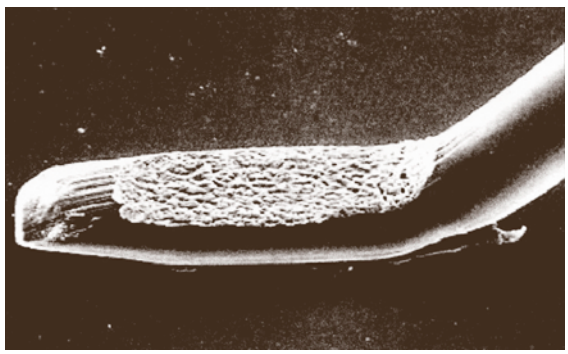


Fig. 7.267 Wedge-joint (after [7.249])

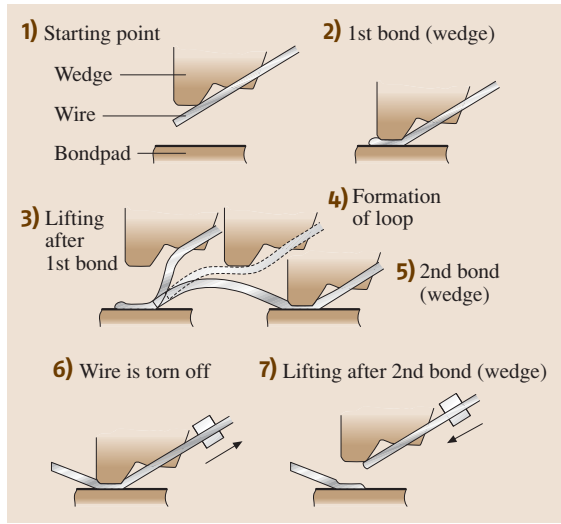


Fig. 7.268 Procedural scheme of wedge–wedge bonding

the beam results in an improved energy input and offers a more appropriate adaptation to the welding job. The parts to be welded have to be exactly positioned as an existing gap can hardly be bridged. Therefore, an easy-to-manufacture perfect fitting accuracy or a sufficient

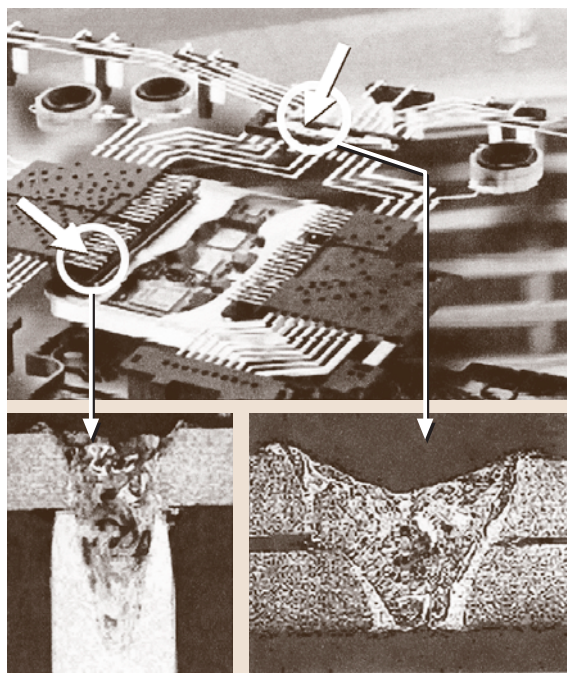


Fig. 7.269 Laser beam welding of electronic parts of a gear control system (after [7.250])

material reserve for melting have to be considered in the designing phase of the parts. An enormous advantage relies on the accessibility of the laser beam to low-lying welding locations. By means of a long welding pulse, limited welding seams are also possible (Fig. 7.269).

Laser beam welding is especially suited for the fabrication of precise mechanically and thermally highly strained components of the ancillary industries in automobile manufacturing and precision mechanics, predominantly in situations where the use of metallic materials exhibiting a reduced weldability is required. In medical technology with its high demands on the surfaces of components as regards cleanliness, sterility, and biocompatibility, precision weldments of pacemaker casings are made without damaging their complex interior. The dental lab uses this technique for the fabrication of bridges and braces. In electrical engineering and electronics it is applied for the manufacture of small parts and electrical contacts. Laser beam welding replaces the previously used soldering techniques in those cases where high thermal demands on precision parts prevail.

Microsoldering Processes

Electronic units made by soldering consist of a combination of single elements that are connected by means of a carrier board. Each of those elements can be a single component, a combination of parts, a highly integrated unit, or an already joined subelement fulfilling an electronic function. The printed circuit board is often part of the mechanical structure and carrier for the conductor tracks connecting the structural elements. In addition, it supports the distribution of heat of the structural elements applied to it. Depending upon the mode of the structural elements applied, the carrier boards are equipped with holes (through hole technology) or without holes. The conductor tracks are situated on one side, on both sides, or in the inside of the boards in the case of multilayer boards [7.252, 253].

The components are placed either on one side of the substrate or on both. The different positions of the elements determine the options for the application of a certain soldering technique (Fig. 7.270).

For microsoldering techniques, SnPb, SnPbCu, and SnPbAg solders are predominantly used. The disposal of used electric and electronic appliances and the restrictions as regards the use of dangerous substances in such devices require a change in manufacturing toward unleaded solders (e.g., solder $Sn_{0.7}Cu$ for wave soldering and solder SnAgCu for the reflow soldering process) [7.254].

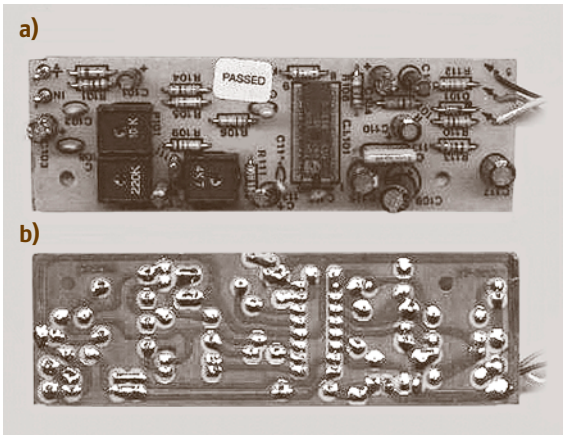


Fig. 7.270a,b Printed circuit board, top **(a)** and bottom side **(b)**

Manual Soldering. Iron soldering is preferably used for small numbers of pieces that have to satisfy extremely high demands as regards the reliability of the soldered joints in the field of air- and spacecraft engineering as well as the repair of microsystems. To reach a consistent soldering temperature, controlled manual soldering devices with corresponding soldering stations are applied. Overheating of the soldered joint and the resulting damage of the board can thus be avoided. However, the quality of the soldered joints depends to a high degree on the skill and experience of the solderer. The industrial manufacture of high numbers of pieces always requires the application of mechanical soldering techniques.

Dip Soldering. In dip soldering, the surface-mounted board is lowered nearly horizontally into the solder bath until the face to be soldered reaches the surface of the liquid solder (Fig. 7.271). After a predetermined dwell time, the board is lifted from the bath surface again. The soldering process is complete.

As a rule, flux application, preheating, soldering, and board transport are not mechanized. Generally, the removal of surface contamination, the lowering of the

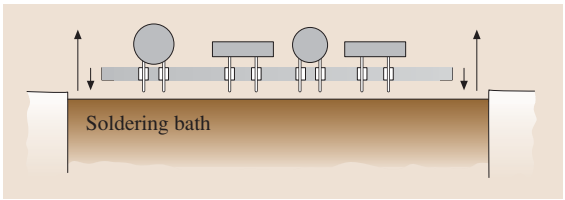


Fig. 7.271 Principle of dip soldering

printed circuit boards until the contact of the face to be soldered with the surface of the liquid solder is reached, and the lifting of the boards after the given dwell time are done automatically.

Drag Soldering. In drag soldering, the surface-mounted board is clamped in a frame with the side to be soldered in the downward position. The frame moved by a chain transport system is lowered into the soldering bath until contact with the liquid solder is reached. Then the board is dragged through the bath over a given distance which corresponds to the soldering time and after that lifted from the bath again (Fig. 7.272). For reaching an optimum soldering quality, the angle of dipping and lifting of the printed circuit board can be changed by adjusting the chain guide.

Drag soldering is a continuous process. Flux application is accomplished by foaming. An oxide slide is used for the removal of the oxide skin which always forms on the surface of the soldering bath prior to the soldering of a new board.

Wave Soldering. In wave soldering, a stationary, regularly recurring soldering wave is generated in the solder bath. The face of the printed circuit board to be soldered is moved in one direction through the wave with a set angle of approx. 7 °C. Frequently, several successively arranged solder waves are applied. Figure 7.273 shows the soldering of printed circuit boards with mixed insertion by means of a wave soldering device. Wave soldering devices operate fully automatically and are controlled by microprocessors.

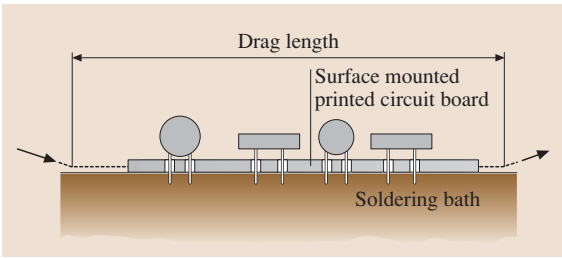


Fig. 7.272 Principle of drag soldering

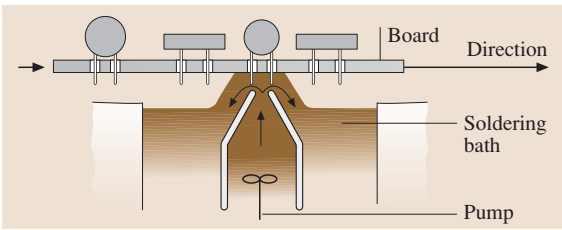


Fig. 7.273 Principle of wave soldering

Reflow Soldering. Surface-mounted devices (SMDs) are miniaturized components that are delivered with pretinned contact faces. They are placed on the pads of the substrate, which are pretinned as well. In the case of small series, the positioning of the components can be done manually; in big series, this is done with the help of pick-and-place as well as positioning machines. The single components are first glued provisionally and then finally integrated into the circuit by the subsequent soldering process. In this case, the reflow soldering process is almost exclusively used.

Nonglued components align themselves by the surface tension of the liquid solder.

In reflow soldering, a previously applied solder is molten during the soldering processes to produce the joint. A further solder application is not required. The single process cycles comprise the manufacture of printed circuit boards with solder pads, the SMDs, and the heat input required for melting the solder. Depending upon the type of SMDs, the quantity of the solder required per soldered joint amounts to 0.5 up to 1.2 mg [7.252].

Reflow soldering comprises the large-scale soldering fabrication and sequential soldering techniques. They can be differentiated by their manner of heat input. In the case of soldering accomplished by the total heating of the workpiece, techniques using hot gas, saturated vapors, and infrared radiation are of major importance as they are well-suited for a continuous soldering process.

For soldering multicontact components and for the repair of soldered joints of surface-mounted components, local heating is deliberately used. Soldering with hot gas (hot air), pulse soldering using a resistance-heated strip or U-strap support (thermode), soldering with infrared radiation, as well as laser beam soldering are of special importance. By using the laser beam also barely accessible soldered joints can be reached.

Fixation and clamping of the components during the soldering process are accomplished by glass plates that can be penetrated by the laser beams. CO_2 lasers and Nd:YAG lasers are preferably applied for soldering.

7.4.7 Microbonding

A few years ago, the market trends for microsystems were primarily determined by silicon-based monolithic systems from the field of information technology.

Efforts to realize functions that are more and more complex and have constantly higher integration densities, however, increased the demand for microsystems that combine different mechanical, optical, fluidic, and electronic components on one common carrier structure. The manufacture of this type of hybrid microsystem demands efficient joining techniques that can be adjusted flexibly to different batch sizes, materials, geometries, and technological requirements. In this connection, microbonding represents a method with many advantages and it is, in some cases, the sole possibility to join dissimilar materials. The term *bonding* stands for the manufacture of a permanent positive substance joint between two joining members by means of adhesion (interfacial bonding) and cohesion (internal strength). The possibility of joining a multitude of different materials and geometries with low temperature is particularly advantageous. In the microrange, the adhesive layer frequently fulfills, apart from mechanical fixing, other additional functions, such as conducting and/or separating of light, electricity, heat, and fluids. In comparison with competitive joining methods, e.g., microwelding and brazing or bonding, the restricted aging and thermal stability exert a negative influence on the creeping behavior of the adhesive layer.

Adhesives

In principle, the same monomer and polymer base materials as are used in the macrorange are applicable.

The demands made on the adhesives are a direct result of the defined application case. They are mainly determined by the joining part materials, the geometrical design, the type and extent of the scheduled duties, and the requested functionality. The properties of the adhesive and/or of the adhesive layer (Fig. 7.274), are important decision criteria.

It is generally valid that, in the microrange, the large ratio that the surface bears to volume may lead to the escape of volatiles from the adhesive layer and thus to a substantial change in the adhesive composition. This effect is increased when the surfaces show strong curvatures. Apart from this, the large surface of small adhesive volumes is characterized by a stronger susceptibility to secondary reactions with the environment, which results in, e.g., the inhibition of the hardening reaction through atmospheric components. For application in the field of microsystem technology, mainly reaction adhesives based on epoxy resin, acrylate, silicon, and polyamide are used. Here, a distinction is made between filled and unfilled adhesives. A summary of the properties of unfilled adhesive systems for the microrange is, among other data, shown in Fig. 7.274.

Selection criteria for suitable adhesives	Characteristic parameters
Material compatibility between adhesive, joining part and surrounding media	• Development of sufficient adhesive forces • No secondary effects through active ions • Physiologically uncritical
Mechanical properties	• Tensile-, pressure-, shear strength • Elasticity modulus, shear modulus • Elongation after fracture • Hardness
Behavior under increased load	• Temperature area of application and temperature behavior – change of the mechanical properties – thermal expansion coefficient • Effects of permanent load – creeping behavior under constant static load – mechanical properties under cyclic load • Aging stability under influence of humidity or corrosive media
Processing requirements and stetting conditions	• Requested storage conditions • Pot life, setting time • Processing time and setting time (pressure)
Functionality of adhesive layer	• Electrical insulation and/or conductivity – specific resistance • Thermal conductivity – thermal transfer resistance • Optical properties – refractive index, transparency

Fig. 7.274 Typical criteria for the selection of suitable adhesive systems

Fillers. Filled adhesives are composed of the same base material as unfilled systems and, as far as quality is concerned, show, therefore, equal material behavior. They contain, in addition, particles that exert a defined influence on the property profile of the adhesive layer. Through the addition of approx. 70 mol % of silicon oxide particles to the epoxy base material, the modulus of elasticity is, for example, increased from 3 to 20 GPa and the thermal expansion coefficient is decreased from 30 to 15 ppm/K [7.255]. The addition of 70 mol % Al_2O_3 increases the heat conductivity from 0.4 to 2.0 W/(m K) and 60–80 mol % Ag, Ni, Au flakes

reduce the conductivity of the set adhesive layer down to values of less than $10^{-5}\ \Omega\,\mathrm{cm}$.

Filled adhesives require special storage. Apart from the premature hardening of the reactive components, the attachment of particles to one another must be avoided. High temperatures intensify this process. During the dispensing process, partly high shear forces that may lead to the separation of matrix and fillers are acting upon the adhesive. The dispensing volume flow must, in this case, be adapted. With decreasing joint geometries the fillers may lead to nonreproducible bonding results as their dimensions are, partly, within the same order of magnitude as those of the adhesive layer. Based on recent research work, nanoscaled fillers are in this connection increasingly part of the adhesives formulas [7.256].

Process Technique

Joining and Setting. The transfer from the macrorange to the microrange entails, apart from the described differences in processing and hardening of the small quantitites of different adhesives, substantially increased demands on the handling precision during joining and setting. The geometrical precision and the rigidity and/or placement force of the system determine the evenness and thickness of the adhesive layer thickness and also the spreading behavior of the adhesive. Typical layer thicknesses in the microrange are between 1 and 20 µm. For component handling, normally pick-and-place systems with precision degrees in the micrometer range are used. Reproducibility and handling precision are, therefore, increased by suitable adjusting structures, Fig. 7.275. Critically positioned adhesive joints are, generally, fixed until the adhesive has set completely and/or at least almost completely.

Surface Treatment. In adhesive bonding, surface treatment is of special importance. Cleaned, degreased, and homogeneous joining part surfaces are a precondition for uniform wetting conditions and for the development of sufficient adhesive forces between adhesive and base material. The adhesive forces which are responsible for adhesion are in the subnanometer range. Treated surfaces show, even after superfinishing, surface roughnesses of more than 25 nm, thus low-viscosity adhesives are frequently connected with improved wetting conditions and a more effective surface. Cleaning and degreasing of the surfaces demands special attention as also supposedly cleaned surfaces may show adsorbate layers which locally change the wetting capability. Due to the naturally small dimensions this leads, in the microrange, to uneven adhesive spreading and to reduced adhesive strengths, Fig. 7.276.

Dispensing and/or Application Technique. For dispensing and applying defined minimum adhesive quantitites, various dispensing techniques have, due to the multitude of different types of adhesives and requirements, been established. A distinction is made between large-surface application methods, such as screening/adhesive printing and point-shaped application methods, such as the pin-transfer technique or the different dispensing techniques by means of which also a linear-shaped adhesive application can be realized. Important properties of the mentioned methods are explained in what follows.

Screening/Adhesive Printing. In screening, the adhesive is pressed through a screen by means of a coating

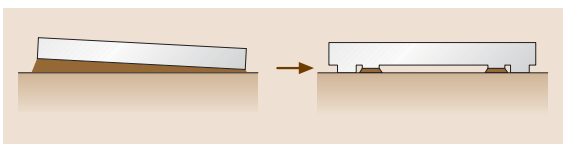


Fig. 7.275 Passive increase of the measuring accuracy through application of spacers

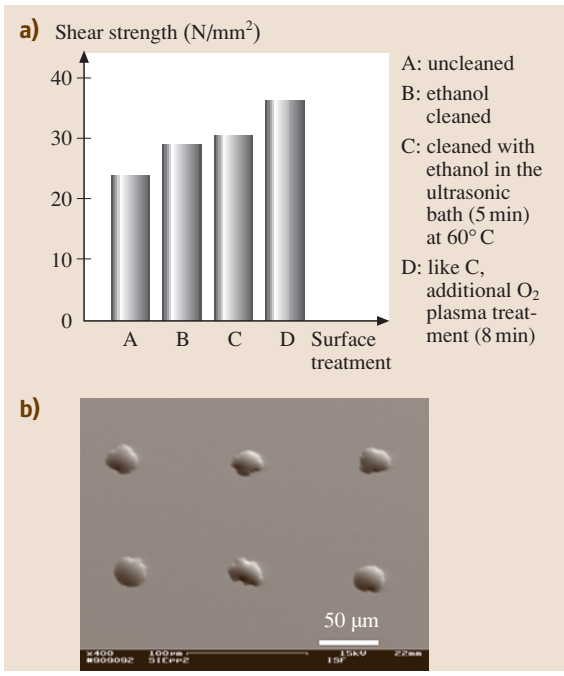


Fig. 7.276a,b Shear strengths on differently cleaned surfaces **(a)**, glass substrate cleaned with ethanol **(b)**

knife which shows recesses in the form of the demanded application geometry. The screen is, in principle, made of steel or nylon and is, by coating knife movement, pressed elastically onto the substrate surface. The application result is determined by the rheological characteristics of the adhesive, the coating knife pressure and speed, and also by the screen geometry and material. The bottom sides of the screen are partly coated in order to prevent threading of the adhesive during the elastic springback of the screen. The stamping method is similar to the screening method. In adhesive printing, the adhesive is, by means of the coating knife, pressed through a metal sheet which is equipped with recesses in the form of the demanded application geometry. In comparison with screening, the adhesive printing method allows the realization of larger printed-layer thicknesses. Both technologies are only suitable for applications on plane surfaces, as, for example, the application of electrically conductive adhesives on printed boards.

Pin-Transfer Technique. For carrying out the pin-transfer technique, a hollow or massive pin is, as a first step, dipped into an adhesive reservoir. Then, a part of the adhering adhesive is transferred to a substrate. The method is simple and characterized by inexpensive design which is easy to parallelize (Fig. 7.277). The technique is suitable for a large viscosity range (100–100 000 mPa s) with, at the same time, large volume spectrum. This includes filled adhesives. The pin-transfer technique, however, is most inflexible as regards a change of the applied volume and of the application geometry. Moreover, the method is characterized by comparatively bad volume exactitudes, a low degree of reproducibility, and high off times.

Dispensing Techniques. The dispensing techniques show, in comparison with the pin-transfer technique and the screening/stamping techniques, substantial advantages as far as flexibility is concerned. The different dispensing techniques allow realization of close to any desired spot or linear-shaped application geometries. With regard to volume and geometry, the individual drops and/or lines are variable over a large range. Typical dispensing volumes reach from the microliter range far into the subnanoliter range. Various dispensing systems which can be differentiated into contact and noncontact systems are used (Fig. 7.278).

In contact dispensing techniques, the adhesive is provided at a capillary by means of a feeding unit

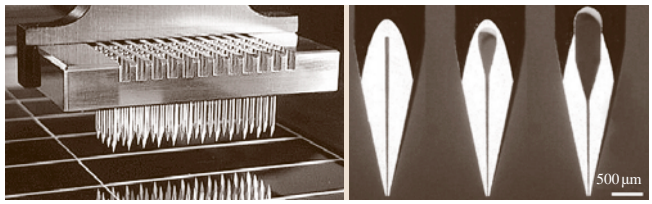


Fig. 7.277 Pin-transfer technique

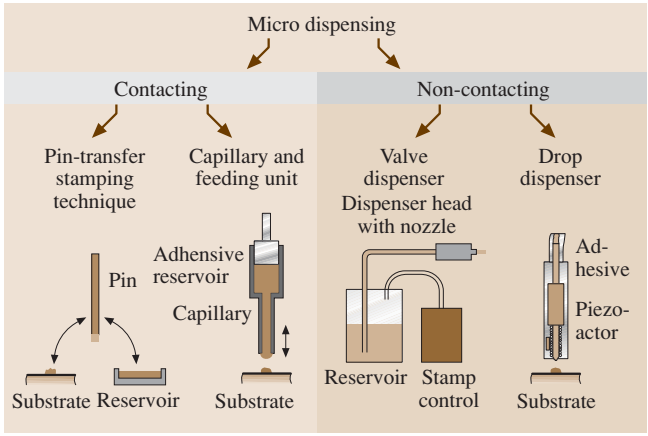


Fig. 7.278 Contact and non-contact micro dispensing

and then transferred to the substrate. The application volume depends mainly on the properties of the adhesive, the capillary, and the substrate [7.257]. The contact methods are characterized by a simple and comparatively inexpensive design. They show advantages particularly in the application of the smallest possible adhesive volumes with, at the same time, large viscosity ranges. The feeding units operate according to different displacement principles. Most frequently applied are piston, screw, peristaltics, and compressed air systems. Figure 7.279 shows a piston dispenser that was developed in the ISF. This unit allows the dispensing of adhesive volumes within the range of a few pl [7.258].

Examples of noncontact dispensing systems are the piezo drop dispenser and the valve dispenser. The piezo dispensers have different designs and operate according to the same principle as ink jet printers. A design which has been established for adhesives application is, for example, a piezoactor, which encloses a glass tube filled with adhesive. The piezoactor is driven and contracted via the voltage, pulse duration, and frequency. As a consequence of the pressure impulse, the adhesive is expelled from the nozzle. The individual drops are within the volume range of 30 and 500 pl. Larger

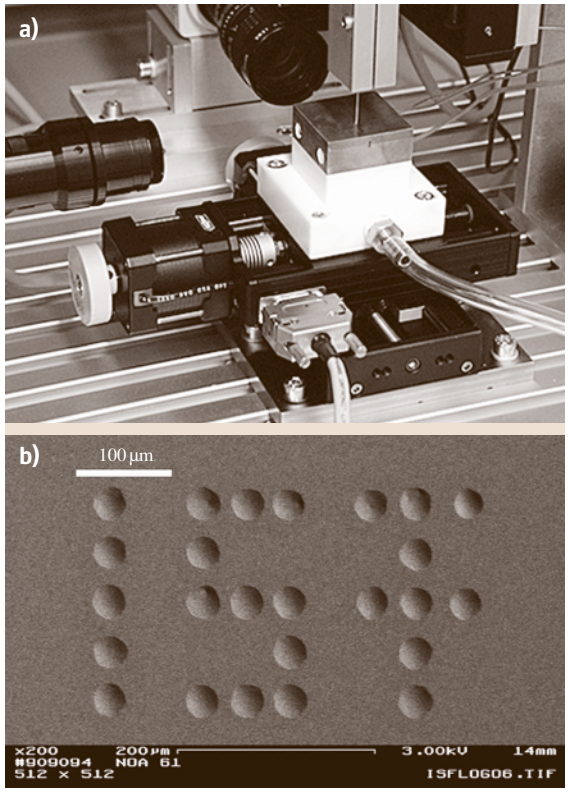


Fig. 7.279a,b Micro piston dispenser (a) and example of dispensing (b)

volumes are realized through multiple dispensing of individual drops at up to 2000 Hz. The volume error is, in this case, lower than 1%. The frequent errors during process interruption and the restricted range of viscosity are disadvantageous. In principle, only unfilled, low-viscosity liquids with viscosities of less than 20 mPa s can be dispensed. With a heated nozzle this range can be increased up to a maximum of 20–10 000 mPa s.

Valve dispensers in general operate with piezo-driven, fast-action needle valves that release an adhesive reservoir that has been under pressure. Modern, heated dispensing systems are capable of dispensing low-viscosity adhesives with viscosities of 70 mPa s up to thixotrope adhesives with viscosities greater than 150 000 mPa s. The minimum adhesive quantities of the individual droplets lie between 2 and 100 nl, where, as a rule, low viscosities make small individual droplets possible. Quantitites of up to 150 individual droplets per second can be dispensed.

A comparison of contact and noncontact dispensing systems shows that the latter have their limits in dispensing very small quantitites and high viscosities. In contrast with contact systems, however, the setting process does not take place. These systems are, therefore, applicable in a much more flexible way and operate mainly independently of the substrate.

7.4.8 Modern Joining Technology – Weld Simulation

Computer simulations are often used to predict useful engineering information regarding weld residual stress and distortion, thus minimizing costly experimental testing and measurements. Simulations are also helpful in the analysis of weld failures, giving insight into which cracking mechanisms are pertinent and which controlling parameters are most important. They are also useful in the development of cracking models, helping to identify which proposed mechanism best fits the observed behavior. Thus, weld simulations have a multitude of uses, but must always be performed in conjunction with experimental tests for verification.

Weld simulations generally require a coupling between two or more models describing thermal, mechanical, or metallurgical behavior. Because welding inherently involves a moving heat source, all things subsequent to welding are related to the weld thermal cycle and its characteristic high temperatures and steep gradients. However, stresses are not all thermal in origin and can be supplemented by stresses associated with solidification and solid-state phase transformations. Reaction stresses, controlled by material resistance and restraining forces, will interact with the local stresses around a moving weld pool to result in global residual stresses upon weld cool down.

A simulation is only as good as the model upon which it is based. It follows therefore that one must have an understanding of the models used and their limitations. Because simulations are based on the coupling of different models for the purpose of predicting weld behavior, the topic of weld simulation must by nature include a discussion of various models as outlined below. Generally, there a two ways to distinguish the various weld simulation procedures. Usually, such simulation tools are divided into weld process simulation and computational weld mechanics.

Taking into account weldability as a component property, being governed by a combination of welding process, construction, and material following DIN 8528 [7.259], a more precise distinction of the various modeling tools is appropriate. Different models in comprehensive simulations, including thermal, structural,

and metallurgical models, interact with each other as shown in Fig. 7.280. Since most simulations are based upon a thermal analysis of weld heat flow and those factors affecting the thermal cycle, this will be addressed first. The subsequent sections then show how structural analyses can be coupled and, finally, it is highlighted how metallurgical models can be incorporated.

Thermal Analysis

Temperature Fields

Heat Conduction. Mathematical modeling of welding heat effects can be carried out analytically or numerically. The well-known anaytical model and associated equations for heat distribution in welding were first developed by Rosenthal [7.260,261]. This model specifically represents heat conduction from a moving heat source in a homogeneous and isotropic continuum, where the field equation is of the differential type of second order

$$\frac{\partial T}{\partial t} = a\left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2}\right), \text{ with } a = \frac{\lambda}{c\rho}. \tag{7.176}$$

It contains the local volume-specific heat capacity $c\rho$ (J/(mm^{3}K)), mass-specific heat capacity c (J/(g K)), density ρ (g/mm^3), as well as the material- and temperature-dependent quantity $\lambda/c\rho$ (thermal diffusivity (mm^2/s)).

Analytical calculation of the temperature condition is by nature quasi-steady state. Further necessary simplifications are made by using idealized geometries and a concentrated heat source in the heat-conductive continuum. *Rykalin* [7.262], for instance, considered idealized basic bodies for determining the temperature distribution during welding, depending on the extent of heat distribution in the component: e.g., single-edge limited body, plane layer in space, plate (plane layer with small thickness), and rod (body with straight or only slightly curved axis and small cross-section).

Due to the development of fast computers and increasingly improved software programs available for PC applications, numerical simulations using a finite-element or finite-difference method have become common. Numerical modeling appears also to be more accurate than simple analytical solutions. For this, the field equation describing heat conduction is transformed into a system of linear equations that can then be solved using linear algebra solutions.

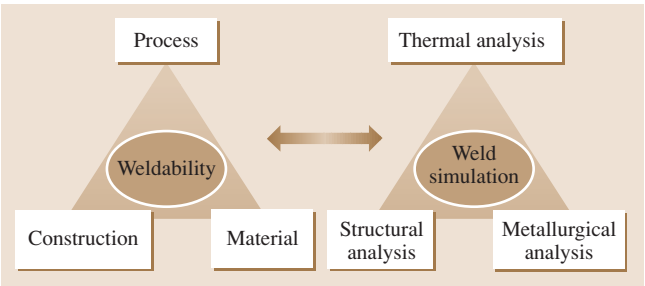


Fig. 7.280 Relation between weldability as a component property, according to DIN 8528 (after [7.259]), and weld simulation

It has to be emphasized that knowledge of the nonlinear material physical properties is important for accurate determination of the location- and time-dependent temperature distribution. For thermal analyses, heat conductivity and arc efficiency are of vital importance [7.263]. However, exact material characteristics for high-temperature applications are available only to a very limited degree [7.264, 265]. For this reason, thermal analytical and numerical calculations of the time-dependent temperature distribution must always be verified by the respective experiments.

Practical and adaptive model networking is essential to achieve economically justifiable computing times. Accordingly, relatively rough model networks or the use of a new style line Gauss heat source are reasonable for numerical temperature field calculations, to achieve agreement between experiment and simulation [7.266–268]. A compilation of typical welding temperature field calculations can be found in Table 7.53.

Specific Heat Sources for Welding. The shape and intensity of the heat source during welding can be completely different from those of other engineering heating processes and, thus, must be incorporated properly into the total thermal model. It must also be mentioned in this context that the temperature field in the component is only dependent on the weld pool up to a distance corresponding approximately to the diameter of the welding heat source. At larger distances the evolving temperature field is determined by the dimensions of the component. For this reason, simplified heat source types have been introduced [7.269], e.g., spot source, linear source, and areal source. In analytical calculations, for example, the spot source can be applied for overlay welding, the linear source for butt welding, and the areal source for friction welding. *Rykalin* [7.262] showed that the heat flow density q for a surface source provides a normal distribution defined by (7.177), and a hemi-

Table 7.53 Examples of analytical and numerical temperature field calculations (after [7.283])

Author(s)	Year	Welding procedure and heat source, respectively	Specific features
Analytical			
Norman et al. [7.264]	1998	Laser welding	Combination of spot and linear source
Kondoh, *Ohji* [7.265]	1998	TIG-welding	Open- and closed-loop process control by comparison of measured and analytical data
Hermans, *den Ouden* [7.266]	1998	Short-circuit gas-shielded metal-arc welding	Calculation of process-specific parameter using specially developed computational algorithms
Suzuki, *Trevisan* [7.267]	2000	Multirun arc welding	Temperature distribution for thin sheet metal
Nguyen et al. [7.268]	1999	Double-ellipsoidal heat source	Analytical solution for a moving double-ellipsoidal high-energy heat source
Kamala, *Goldak* [7.270]	1993	Double-ellipsoidal heat source	Error assessment of analytical 2-D-models
Eagar, *Tsai* [7.271]	1983	Gauss heat source	Determination of the HAZ-geometry
Kasuya, *Yurioka* [7.272]	1991	Quasi-stationary, instantaneously active and non-stationary heat source	Determination of the HAZ-geometry and of the $\Delta t_{8/5}$-times
Jeong, *Cho* [7.273]	1997	TIG-welding and tubular cored arc welding	HAZ- and temperature field calculation with the help of a 2-D-normal distribution Gauss heat source
Sudnik et al. [7.274]	2001	Metal active-gas welding	Mathematical model of a heat source during metal active-gas welding
Nguyen et al. [7.275]	2001	Hybrid double-ellipsoidal heat source	Superposition of semi-elliptical heat sources for better approximation to the real shape of the melt
Numerical			
Bonifaz [7.276]	2000	Arc welding	Comparison between experiment and calculations
Murugan et al. [7.277]	2000	Manual metal-arc welding	Multirun welding
Gu et al. [7.278]	1993	Arc welding	Comparison between static temperature field calculations with transformed Euler's formulation and calculations with Lagrange formulation
Murugan et al. [7.279]	1999	Manual metal-arc welding (overlay welding)	Temperature field calculation, microstructure hardness
Little, *Kamtekar* [7.280]	1998	TIG-welding	Investigations into the influence of individual temperature-dependent parameters on the simulation results
Cai et al. [7.281]	2001	Line Gauss source	Efficiency improvement with application of the line Gauss source
Zhou et al. [7.282]	2003	Two-side TIG-MIG-welding	Temperature field calculation, geometry of the melt

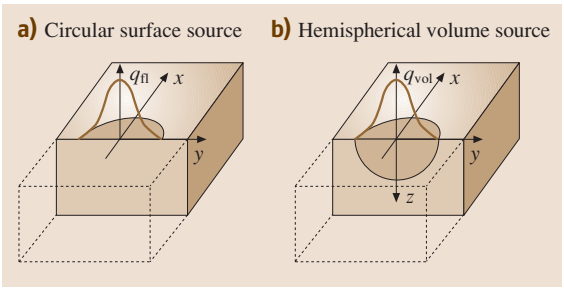


Fig. 7.281a,b View of welding heat sources: Circular surface source **(a)** and hemispherical volume source **(b)**, as defined by *Radaj* (after [7.269])

spherical volume source with normal heat flow density distribution is defined by (7.178) (both are depicted in Fig. 7.281)

$$q_{fl}(x, y) = \frac{3q}{\pi r_{0.05}^2} e^{\left(-\frac{3(x^2+y^2)}{r_{0.05}^2}\right)}, \quad (7.177)$$

$$q_{vol}(x, y, z) = \frac{6\sqrt{3}q}{\pi\sqrt{\pi} r_{0.05}^3} e^{\left(-\frac{3(x^2+y^2+z^2)}{r_{0.05}^2}\right)}. \quad (7.178)$$

There also exist heat source equations that account more appropriately for specific welding processes, like the nonsymmetric-ellipsoidal volume heat source developed by *Goldak* [7.263] shown in Fig. 7.282. This heat source is characterized by front and rear quarter ellipsoids, which can be defined independently of each other. This model takes account of a realistic nonsymmetric heat source density distribution, (7.179), which is always greater in the front semiaxis $x_{0.05f}$ than in the rear semiaxis $x_{0.05r}$. A further advantage is to be seen in the fact that the thermal output figures (f_f and f_r) may be assumed to be different in the front and in the rear ellipsoid

$$q_{volf,r} = f_{f,r} \frac{6\sqrt{3}q}{\pi\sqrt{\pi} x_{0.05f,r} y_{0.05} z_{0.05}} \times e^{\left\{-3\left[\left(\frac{x}{x_{0.05f,r}}\right)^2+\left(\frac{y}{y_{0.05}}\right)^2+\left(\frac{z}{z_{0.05}}\right)^2\right]\right\}}. \quad (7.179)$$

In order to support the modeling of different thermal weld cycles and their respective heat source types, numerous temperature field calculations are listed in Table 7.53.

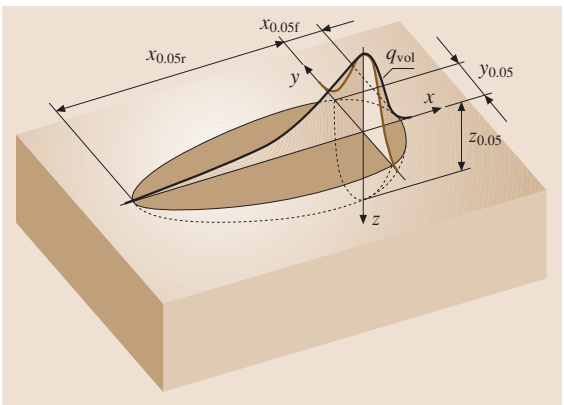


Fig. 7.282 Unsymmetric-ellipsoidal volume heat source with normal distribution as defined by *Goldak* et al. (after [7.263])

Thermomechanical Analysis

Calculation of Welding Stresses and Strains. The thermal heating and cooling cycle during welding causes stresses and strains in the longitudinal, transverse and thickness directions of the joint, which are most relevant

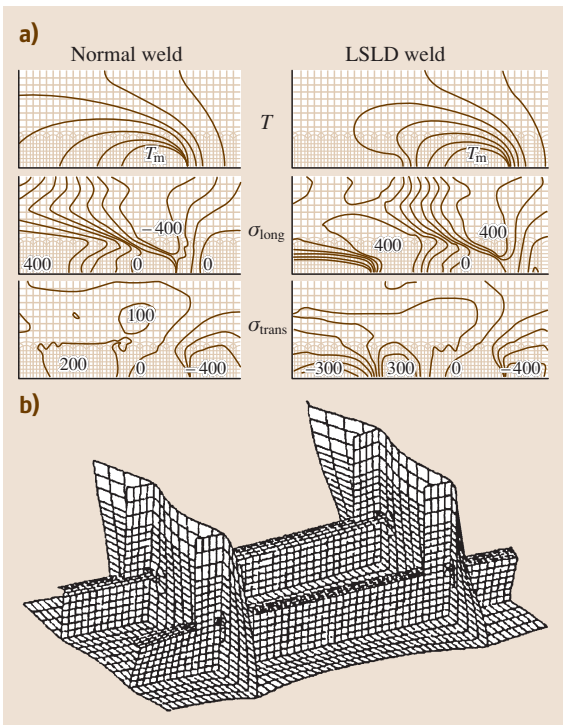


Fig. 7.283a,b Thermo-mechanically calculated distributions of residual stresses and strains in welded joints: **(a)** Temperature and stress contours (after [7.284]), **(b)** welding deformation in test model (after [7.285])

Table 7.54 Approaches to transverse shrinkage assessment of single-pass butt welds (after [7.297])

Method proposed by	Heat input q_s, $S_y = f(q_S)$	Weld cross section A_0, $S_y = f(A_0)$
Okerblom, Michailov et al. [7.288–290]	$S_y = 2A_y \frac{\alpha}{c\rho} \frac{q_s}{h}$ (7.180)	
Wörtmann and Mohr [7.291]	$S_y = 2A_y \frac{\alpha}{c} \frac{A_0 q_m}{h}$ (7.181)	
Malisius [7.292]		$S_y = x \frac{A_0}{h} + 0.0121b$ (7.182)
Satoh, Ueda et al. [7.293–295]	$S_y = h_{SG}\alpha \sqrt{\frac{T_S q_m \tan\frac{\phi}{2}}{c}}$ (7.183)	
Watanabe and Satoh [7.296]	$S_y = C_1 \frac{A_0}{h^2} \ln \frac{m_{LE}}{m_{LE_1}} + C_2 \sqrt{\frac{A_0}{h^2}}$ (7.184)	
Matsui [7.297]	$S_y = \frac{q_s}{c\rho h} \mathrm{erf}(f)$ (7.185)	
Gilde [7.298]	$S_y = 0.24 \frac{\alpha}{c\rho} \frac{q_s}{h}$ (7.186)	
Capel [7.299]	$S_y = 17.4 \frac{q_s}{h}$ (7.187)	
Spraragen and Ettinger [7.300]		$S_y \approx 0.25 \frac{A_0}{h}$ (7.188)
Richter and Georgi [7.301]		$S_y = n \frac{A_0}{h} + m$ (7.189)

for the fabrication process and, in particular, for the in-service behavior of the respective component. It is thus a major goal of analytical and numerical simulations to assess such stresses and strains, in order to predict buckling and load-bearing capacities. Such stresses and strains should thus be determined in the design phase, when considering the life cycle of a structure or a component. Significant work has been initiated to improve analytical and, in particular, numerical calculations of weld stresses and distortions, and it has to be

expected that such procedures are increasingly applied during the design of welded constructions, as shown in Fig. 7.283a. In this context, stresses and strains transverse to the welding direction will be of major interest in the future, because with the increasing application of high-strength materials, modern lightweight structures will be designed with more stiffeners transverse to the joint direction. Finite-element analyses [7.286] have already demonstrated that transverse shrinkage is, for the most part, caused by thermal contraction of the constructional members adjacent to the weld and that shrinkage of the weld itself only accounts for approx. 10% of the total shrinkage. A summary of simple formulas [7.287] is given in Table 7.54.

More and more, such thermomechanical calculations are connected to respective metallurgical modeling like the investigations of the effects of low-temperature phase transformations on the residual stress distribution, which will be described in the next section. Although there is a general tendency away from the thermomechanical calculations of welding stresses and strains at simple butt joints toward more complex structures (Fig. 7.283b), only a few attempts have been made to consider the thermomechanical interaction that a specific joint may be subjected to in a real construction. Such calculations are applied to real components only to a limited extent, usually due to extremely high computation time and storage capacities, which are required for accurate computational weld mechanics.

An alternative approach to transferring the thermomechanical interaction between the component and a weld joint, from simulations to reality and vice versa, is given by the concept of the intensity of restraint. This approach considers the transverse stresses and strains in a welded joint, which significantly increase the more the shrinkage is hindered by respective assembly parts. In order to quantify such shrinkage restraint, *Satoh* et al. [7.293–295, 302] introduced the restraint intensity principle shown in Fig. 7.284. It is based on the consideration that the suppressed shrinkage transverse to the welding direction S_y is taken up by the weld metal and base metal. Under the assumption that the base metal behaves elastically and that the weld might also plastify to a limited extent, the intensity of restraint of the weld by the surrounding construction is given by the reaction force F divided by that part of the suppressed shrinkage that is taken up by the base metal S_{yB}. The intensity of restraint thus, in principle, represents a spring constant of the construction part transverse to the welding direc-

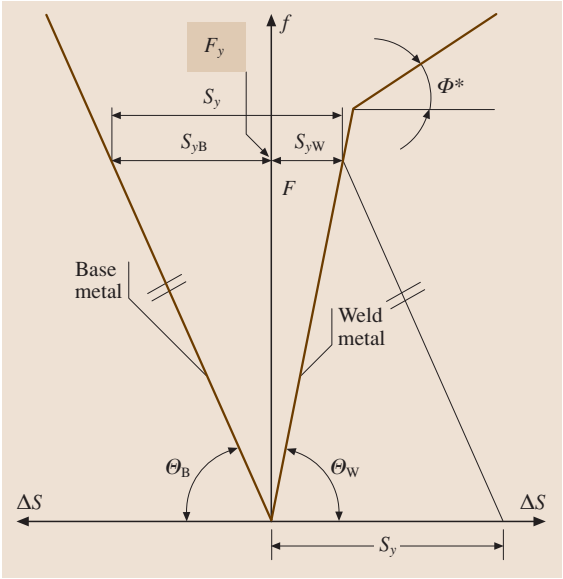


Fig. 7.284 Restraint diagram for the intensity of restraint concept (after [7.293–295, 302])

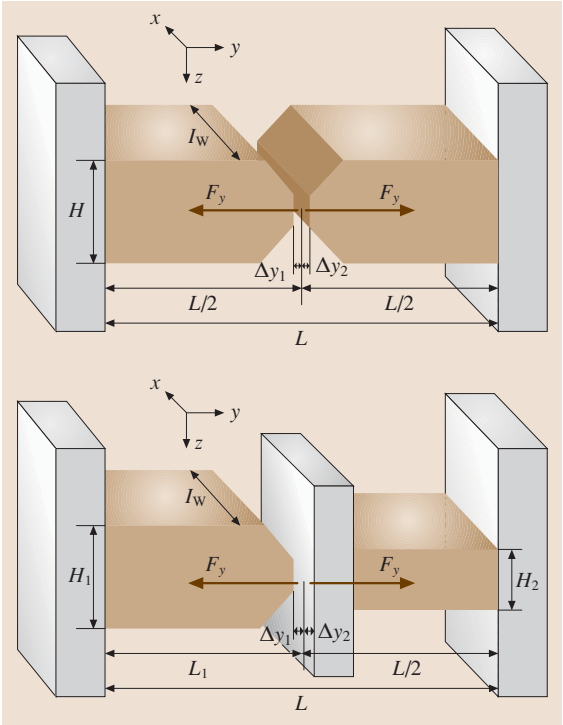


Fig. 7.285 Restraint and reaction forces at symmetric and at asymmetric joints (after [7.303])

tion, normalized to the weld metal length l_W as depicted in Fig. 7.285 [7.293–295]

$$R_{F_y} = \frac{F}{s l_W}\,, \quad R_{F_y} = \frac{F}{2\Delta_y l_W}\,,$$
$$R_{F_y} = \frac{F}{(\Delta_{y1} + \Delta_{y2}) l_W}\,,$$
$$R_{F_{y,tot}} = \frac{1}{R_{F_{y,1}}} + \frac{1}{R_{F_{y,2}}} + \frac{1}{R_{F_{y,3}}}\,. \qquad (7.190)$$

Since the base material behaves elastically, the transverse shrinkage restraint of a specific weld joint can be evaluated by application of a respective force transverse to the welding direction and by measurement or calculation of the respective root gap displacement Δy. More conveniently, the intensity of restraint can be calculated with finite-element analysis during the design phase, where the transverse residual stresses can be assessed, knowing the transverse shrinkage restraint [7.303, 304]. Such calculations confirm that the intensity of restraint for T- and orbital joints can become significantly higher than that of similarly dimensioned butt joints (Fig. 7.287). Dividing the region around a restrained component weld into a near and far field [7.305], it has been found that the total restraint of a joint results from a combined contribution of weld edge preparation, $R_{F_{y,1}}$, plate dimensions, $R_{F_{y,2}}$, and surrounding assembly, $R_{F_{y,3}}$, respectively, as illustrated in Fig. 7.286 and defined in (7.190) [7.303].

For numerical thermomechanical analyses using finite elements, the temperature field analysis and

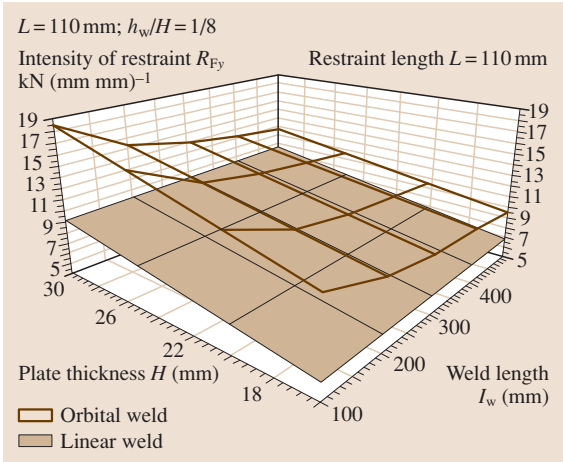


Fig. 7.286 Restraint intensity for linear and orbital welds as a function of plate thickness and weld length (after [7.303])

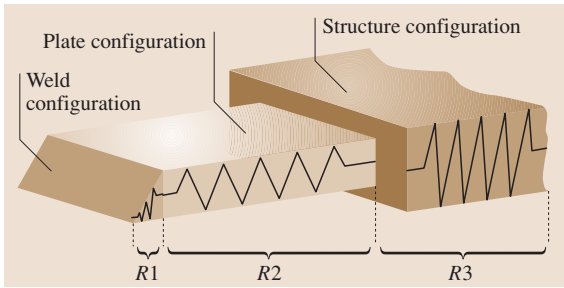


Fig. 7.287 Restraint intensity subdivision model (after [7.283])

the structural analysis can be linked in various ways [7.283]. Usually, such simulations are carried out in an indirect coupled or in a decoupled modus. In a completely decoupled calculation, the temperature data determined by thermodynamic analysis are transferred to the thermomechanical model as time- and location-dependent input data for computation. For such a procedure, the time step intervals have to be identical in the temperature field analysis and the subsequent structural analysis.

A factor of crucial importance for numerical analyses of the temperature-dependent stress–strain distribution in welds is the availability of appropriate material properties. Depending on the computation objective, various accuracy categories have been established in welding engineering (Table 7.55).

As already mentioned with respect to restraint intensity, the accurate transfer of the real component geometry to the finite element code is additionally important in numerical calculation of welding-specific residual stresses and component distortion. The model depth (degree of detailing) can thus be structured in the following ways [7.283]:

2-D Models. Numerical calculation of welding-specific stress conditions using 2-D FE models constitutes a strong simplification of real conditions. With a very low stress gradient in the thickness direction (thin sheet metal), the existent plane stress condition (PSC) can be represented with the help of a 2-D model.

Combined Models. Combined models of the area near the weld usually consist of 3-D networks which allow statements to be made regarding stress as well as component distortion in all coordinate directions. Component areas at a greater distance from the weld, in which a stationary stress distribution prevails, are two-dimensionally networked (3-D–2-D model). Furthermore, the influence of the component/surrounding

Simulation accuracy	Computation objectives
Reduced accuracy level simulation	Simple, fast models
Basic simulation	Calculation of residual stress and distortion for simple geometries Neglect of the material behaviour in the high-temperature range
Standard simulation	Calculation of residual stress and deformation for complex geometries Calculation of transient stress and deformation
Accurate simulation	Calculation of residual stress and deformation for complex geometries taking account of the microstructure
Very accurate simulation	Extended detailed simulation, e.g. hot cracking

Table 7.55 Accuracy categories for welding simulation (after Lindgren [7.283])

structure can be represented by 1-D rod or spring elements (3-D–1D model).

3-D Models. For detailed calculation of the residual stress and distortion behavior of welded components, 3-D models are now used for the most part. In order to optimize the computation time, super elements and/or (statically, dynamically) adaptive networks are employed in the procedure. Super elements allow it to integrate the properties of complete workpiece regions and to reduce at the same time the number of degrees of freedom. The idea behind adaptive networking is a network refinement of the areas with expected high temperature and stress gradients. The welding residual stress histories of multipass welds change with specimen thickness, heat input, and phase transformation effects.

Numerical calculations of the thermal stress–strain still have some limitations, as for instance:

- Exact simulation of the specific volume increase in the case of phase transformations
- Limited thermophysical and mechanical properties in the high-temperature range (200–400 °C)
- Incorporation of exact welding TTT diagrams
- Application of elastoplastic principles
- Calculation convergence at high strains and accuracy of results

Goldak [7.306] turned these unsolved tasks into ten major challenges for future thermomechanical simulations, and it can only be emphasized that computational thermomechanical results should generally be confirmed experimentally. Figure 7.288 illustrates the significance

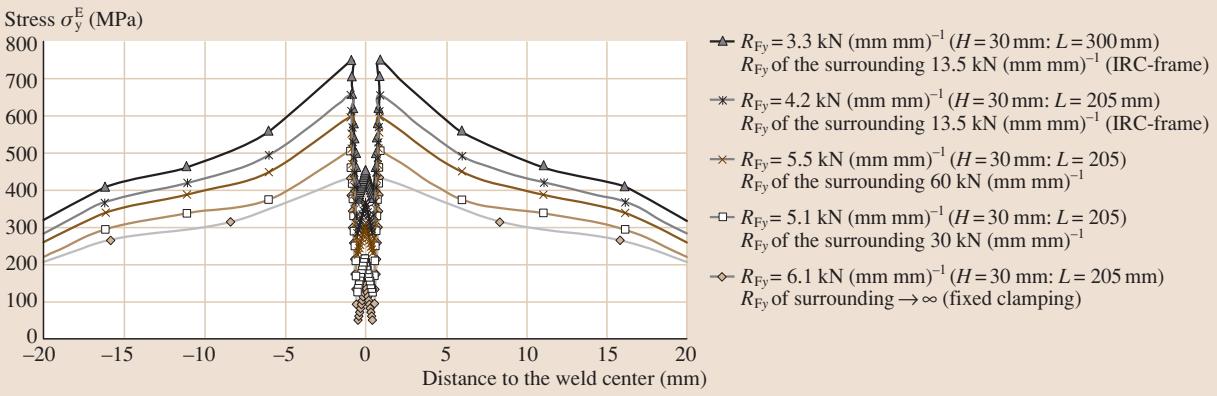


Fig. 7.288 Reaction stresses transverse to the welding direction dependent on the intensity of restraint R_{F_y} (after [7.283])

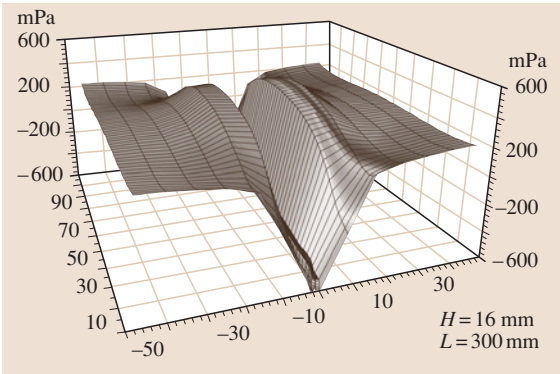


Fig. 7.289 Distribution of transverse reaction stresses in a V-beveled butt joint at $R_{F_y} = 2.1\,\mathrm{kN(mm\,mm)^{-1}}$ (after [7.283])

of an increasing restraint intensity on the reaction stresses transverse to the welding direction; thus the experimentally validated numerical simulations should be carried out with the best possible specimen analogy and representation of the actual shrinkage restraint of the real surrounding structure. This can most conveniently be accomplished by respective component weld tests or full-scale tests. This means the numerical simulation and its experimental validation have to be carried out considering very carefully the intensity of restraint for an accurate transfer of the simulations. Under such conditions, the results of thermomechanical numerical simulations can be transferred into reality and vice versa. Three-dimensional diagrams of the stresses and strains, as shown in Fig. 7.289, can significantly support design optimization of welded components to avoid high local stresses and strains introduced during welding, and to provide respective larger service load capacities.

Weld Pool Fluid Mechanics

Weld Pool Shape. The shape of the weld pool in arc welding is of significant practical importance and the depth of penetration is usually of particular interest. Most engineering applications require that a certain penetration be achieved to meet specific structural and quality assurance requirements (e.g., to avoid lack of fusion). Considering heat conduction alone, and assuming a point heat source (e.g., the Rosenthal equation discussed earlier), the pool cross-section is predicted to be a hemisphere with a 0.5 depth-to-width (D/W) ratio. That this shape is seldom achieved (e.g., normally, D/W ≈ 0.4 for GTAW) can be attributed to two things: (1) a broad, Gaussian heat source (see discussion on heat flow above) and (2) convective heat flow, transferring heat from directly under the arc to outer regions in the pool.

Numerous factors control the convective flow of liquid in a weld pool including arc force, buoyancy force, magnetic force (Lorentz), and surface tension (Marangoni). The arc and buoyancy forces are small relative to Lorentz and Marangoni forces, and are often neglected. The Lorentz force dominates at high currents (> 100 A), whereas the Marangoni force may dominate at low currents. The Lorentz force causes fluid flow towards the pool center and downward, regardless of current polarity, and thus contributes to improved penetration. The magnitude of this force varies directly with current raised to the fourth power (I^4). Because of this strong dependence on current, this is the process parameter normally used to control penetration. It is also commonly observed in empirical equations used to predict penetration (P), e.g., the Jackson equation given below [7.307]

$$P = k\left(\frac{I^4}{SV^2}\right)^{1/3}, \qquad (7.191)$$

where k is a constant, S is travel speed, and V is arc voltage. It should be pointed out from this equation that increases in voltage or travel speed normally result in diminished penetration. At higher voltage the arc spreads out, distributing heat over a larger area. At higher travel speeds, the heat input per length of weld is reduced.

The Marangoni force is related to surface-tension temperature-gradient-driven fluid flow, where flow at the pool surface progresses from regions of low surface tension to regions of high surface tension [7.308]. It is assumed that the center of the pool, directly under the arc, is hotter than at the pool perimeter. How surface tension varies across this temperature gradient depends upon the presence of surface active elements, primarily sulfur and oxygen. For alloys with $\mathrm{d}\gamma/\mathrm{d}T > 0$ (e.g., steel with high sulfur: $S > 300$ ppm), the center of the pool will have the highest surface tension, thus the flow will be toward the pool center and down, enhancing penetration. For alloys with $\mathrm{d}\gamma/\mathrm{d}T < 0$ (e.g., steel with low sulfur: $S < 30$ ppm), flow will be from the pool center to the pool periphery, thus inhibiting penetration. This unusual behavior caused considerable concern when ultra-high-purity, vacuum-arc-remelted stainless steels were first introduced in the 1970s; they were considered unweldable because they could not be penetrated.

Numerical simulations have been used effectively to accurately predict weld pool shape, accounting for

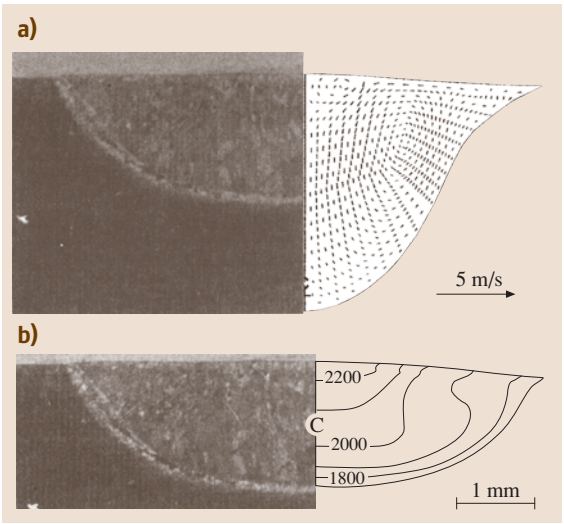


Fig. 7.290a,b Cross-section of a weld made on AISI 304 stainless steel (0.005 wt % S) compared with two different predicted weld pool shapes assuming **(a)** laminar flow and **(b)** turbulent flow (after [7.313])

the forces outlined above [7.309–312]. In more recent work, a distinction has been made between conditions giving laminar versus turbulent flow, permitting more accurate pool shape predictions [7.313], an example of which is given in Fig. 7.290.

In joining processes where the heat source power density is much higher than for normal arc welding (e.g., plasma arc, EB, and laser beam processes), considerable vaporization of the weld pool may occur allowing formation of a keyhole (i. e., a vapor cavity extending thru-thickness). When operated in the keyhole mode, these high-power density processes permit the welding or cutting of thick material in a single pass. This offers considerable savings in terms of labor and consumable costs, plus these joints typically have better mechanical properties due to smaller weld cross-sections and HAZs, and fewer defects normally attributed to multiple pass welding. Popular today is a hybrid combination of processes (e.g., plasma arc plus laser) which combine the keyhole-forming ability of lasers with the better seam tracking, penetration, and stability characteristics of plasma [7.314]. Stability is a topic of special interest in modeling keyhole welding, due to the tendency for molten metal at the trailing edge to collapse the keyhole, resulting in spiking (i. e., variable penetration) and entrapment of large pores. One can consider the force balance associated with a stable keyhole [7.315], where the pressure of the gas flow must be equal to or greater than the surface tension pressure plus the hydrostatic head

$$\frac{\rho_g v^2}{2} \geq \frac{\gamma}{r} + \rho_L g t \,, \tag{7.192}$$

where ρ_g and ρ_L are densities of the gas and liquid metal, respectively, v is the gas velocity, γ is the surface tension, r is the radius of curvature of the liquid surface, g is the gravitational acceleration, and t is the plate thickness. This process has been numerically modeled to show keyhole formation and collapse [7.316].

Weld Pool Oscillation. Weld pools are observed to oscillate at a characteristic, natural frequency dependent upon pool size, geometry, material, and mode of oscillation. Fundamentally, the natural frequency of an oscillating molten droplet is described by the relationship [7.317]

$$f = N\sqrt{\frac{\gamma}{\rho V}} \,, \tag{7.193}$$

where N is a mode-dependent constant, γ is the surface tension, ρ is the density, and V is the volume. If simple pool shapes are assumed (hemisphere for partial penetration, cylinder for full penetration), the volume term above can be expressed in terms of pool diameter (D). The following expressions evolve, with constants determined from experimental data for plain carbon steel welds on a 4-mm plate [7.318]

Mode 1 (partial penetration, peak–valley)

$$f = 2.030 D^{-3/2} \,; \tag{7.194}$$

Mode 2 (partial penetration, slosh)

$$f = 967 D^{-3/2} \,; \tag{7.195}$$

Mode 3 (full penetration)

$$f = 260 D^{-1} \,. \tag{7.196}$$

Use has been made of natural frequency measurements in order to monitor and control weld penetration. When a weld goes from partial to full penetration, its mode of oscillation changes (e.g., Mode 1 to Mode 3). This typically results in a large decrease in natural frequency and a higher amplitude of oscillation. This change can be detected from pool oscillation measurements and can be used for computer feedback control of penetration, where the arc current is normally adjusted to maintain penetration [7.319, 320]. Pool oscillation can be measured indirectly from arc voltage or light intensity measurements, with signals processed using a fast Fourier transform (FFT) to identify natural frequencies [7.321–324].

Thermo-Mechanical-Metallurgical Analysis

Diffusion Analysis. Most of the metallurgical phenomena related to welding are dependent on the kinetic process of diffusion, which can be defined as an atomic transport of matter in a metal matrix [7.325, 326]. In welding, diffusion normally occurs under conditions of high stress and temperature gradients. Because most diffusion processes can quite conveniently be numerically simulated, modeling of diffusion will be discussed ahead of the other simulations related to weld metallurgy. With regard to welding applications, diffusion is often used to investigate the transport of interstitials, e.g., monatomic hydrogen or nitrogen moving through a homogeneous metal lattice.

Under ideal conditions, i. e., by exclusion of additional effects, the flux of the interstitial atoms passing through a unit plane is proportional to the concentration gradient, where the proportionality constant is the diffusion coefficient. This is Fick's first law, which in the 1-D case can be written as

$$J = -D_x \left(\frac{\partial C}{\partial x} \right) , \tag{7.197}$$

where J is the flux of the substance passing through the specific plane and C is the concentration of the substance. The diffusion coefficient varies with temperature and can be described by an Arrhenius relationship as

$$D_x = D_0 \exp \left(-\frac{E_A}{RT} \right) , \tag{7.198}$$

where D_0 is the diffusion coefficient and E_A is the activation energy. Usually, the continuity equation of the conservation of matter is true under such conditions. This means that the time-dependent change in concentration equals the divergence of the flux, and thus for the 1-D case

$$\frac{\partial C}{\partial t} = -\frac{\partial J}{\partial x} . \tag{7.199}$$

The combination of (7.197) and (7.199) gives Fick's second law in the 1-D version [7.325]

$$\frac{\partial C}{\partial t} = D_x \left(\frac{\partial^2 C}{\partial x^2} \right) \text{ and } \frac{\partial C}{\partial t} = \nabla (D \nabla C) ,$$

for the multidimensional version. (7.200)

Such diffusion processes can be numerically modeled using the thermal module of commercial finite-element programs by assignment of the heat conductivity K to the so-called effective diffusion coefficient D_{eff} ($K \Leftrightarrow D_{\text{eff}}$) and by setting the specific heat C_p and the density ρ to unit one ($C_p = 1$, $\rho = 1$) [7.327–329]. Modeling by Fick's laws can be applied in good agreement with experimental results not only for the diffusion of an interstitial like hydrogen, nitrogen, or carbon in a metal lattice, but also for the diffusion of such atoms in homogeneous microstructures, if they are bound at specific sites, but they can all still be activated at the respective thermal conditions. The latter process is generally regarded as reversible trapping.

In the case of irreversible trapping, part of the hydrogen will be kept tenaciously so that this fraction will no longer take part in the diffusion process. Beginning with McNabb and Foster, quite a number of researchers have developed models for such trapping processes, in particular for hydrogen, in the past 50 years [7.327–329], but still with a lack of consistency in experimental results.

In particular, hydrogen has often been reported to be accumulated in weld regions with high residual stresses and strains. Such attraction of hydrogen in crack-susceptible regions is considered to be based on diffusion-enhancing effects, like hydrogen transport with moving dislocations, etc.. Such diffusion enhancing by mechanical stressing or straining of the microstructure can be modeled by inserting an additional potential field in the 3-D version of Fick's second law, which has been investigated extensively by Sofronis et al. [7.330]. However, such numerical analyses have not yet been sufficiently verified by experimental results.

It has been discovered in the past 10 years that modeling hydrogen diffusion in weld microstructures, in particular for carbon, martensitic, and ferritic steels, can be carried out quite consistently by experimentally validated numerical analyses based on Fick's laws [7.327–329]. This means that numerical simulation of a hydrogen concentration profile is possible for the most hydrogen-susceptible weld microstructures. The most important results of such modeling procedures are the development of geometrical hydrogen distribution versus time and the determination of respective removal heat treatments. It can only be emphasized that for such calculations the correct temperature diffusion coefficients (Fig. 7.291) for the particular weld microstructures and thermal cycle have to be inserted into such numerical simulations [7.328]. More recently, the numerical calculation of the geometrical and thermal hydrogen distribution in weld microstructures has been extended by simulation of respective crack initiation and propagation [7.329].

Solid-State Transformations

Welding Specific Transformation Behavior. Knowledge of the thermomechanical relationships between material (welding-specific microstructure transformations) and structure (external structural shrinkage restraint) is becoming of great interest to welding fabrication. Deliberate load-relieving reactions, resulting from the phase transformation behavior of a particular base or filler metal combination, are gaining increasing importance with regard to their influence on the global structural load during fabrication and service. Effects of filler materials with low transformation temperatures on reducing the reaction forces and respective stresses in restrained joints have been reported for high-strength structural steels [7.331] as well as for supermartensitic stainless steels (Fig. 7.292).

The current simulation tasks thus include, above all, studies of the effects of microstructural transformation (martensitic transformation) on various base and filler material combinations (mismatch) as well as studies of the effects of varied heat input during single-pass and multipass welding on the structural load level and distribution of a welded component. The observed significant effects of low-temperature phase transformations on the residual stress distribution were confirmed by *Karkhin* et al. [7.332] by numerical simulations of laser welds (Fig. 7.293). As also stated, the accuracy of such simulations is very sensitive to the measurement of steel properties in the phase transformation range, which thus represent the most important input data for the numerical analyses. In addition, it has to be pointed out that it is essential in the numerical calculation to also take into account the influence of latent heat (transformation enthalpy) on the temperature field which, depending on the weld volume, is basic to an accurate analysis of the global structural load as a result of shrinkage restraint and phase transformation. As concerns higher- and ultra-high-strength materials, the aspect of a load-dependent accelerated phase transformation and of a correspondingly induced transformation plasticity (load-dependent transformation behavior) has additionally to be considered, specifically at high weld stiffness with potential plastic predeformations.

Investigations of the phase transformation behavior in freely shrinking welds generally entail highly simplified assumptions of residual stress levels and distributions in welded structures, e.g., general existence and equal distribution of tensile residual stresses in the amount of the yield strength (or of a specific percentage) and thus, still lead to too conservative designs [7.333]. Thus it is again emphasized that one must clarify the interaction of basic microstructure-specific processes in a real welded structure as well as provide quantitatively correct numerical and experimental coverage of the timely and locally nonstationary metallurgical and thermomechanical processes in the weld. This must be done in consideration of the influence of defined external geometric and structural shrinkage restraints.

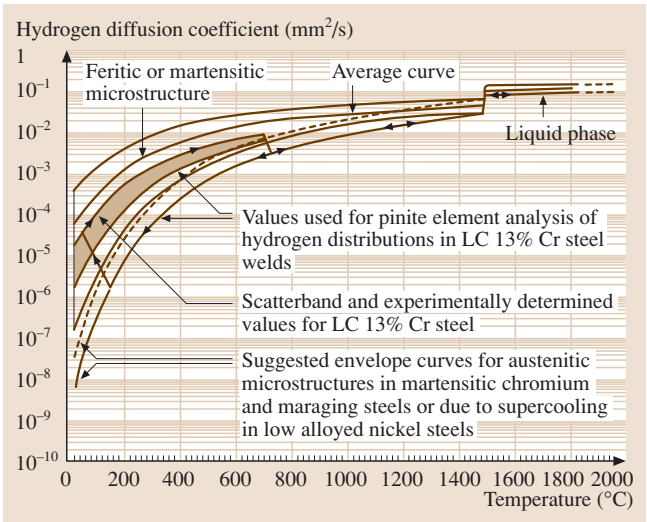


Fig. 7.291 Temperature dependent hydrogen diffusion coefficient in microalloyed structural steels (compilation of respective literature data into a scatterband) (after [7.328])

Hydrogen-Assisted Cold Cracking. Hydrogen-assisted cracking (HAC) of welded components has to be distinguished as two separate phenomena: hydrogen-assisted cold cracking (HACC) and hydrogen-assisted stress corrosion cracking (HASCC). The first appears predominantly during or in a limited time after the fabrication process, while the latter occurs during service of welded components, usually significantly reducing its life time. From the engineering standpoint, both phenomena most conveniently and more macroscopically can be regarded as an interaction between the local hydrogen concentration, mechanical load, and weld microstructure.

HACC and HASCC can then be considered as very similar failure phenomena with their differences only in the source of hydrogen. Regarding HACC, hydrogen is primarily transferred to the liquid weld pool after dissociation in the arc. It usually is introduced into the welding process by moisture on the base and filler materials, leaking cooling systems, and, sometimes, by nonremoved primers and coatings before welding. Regarding HASCC, hydrogen is taken up from a corrosive

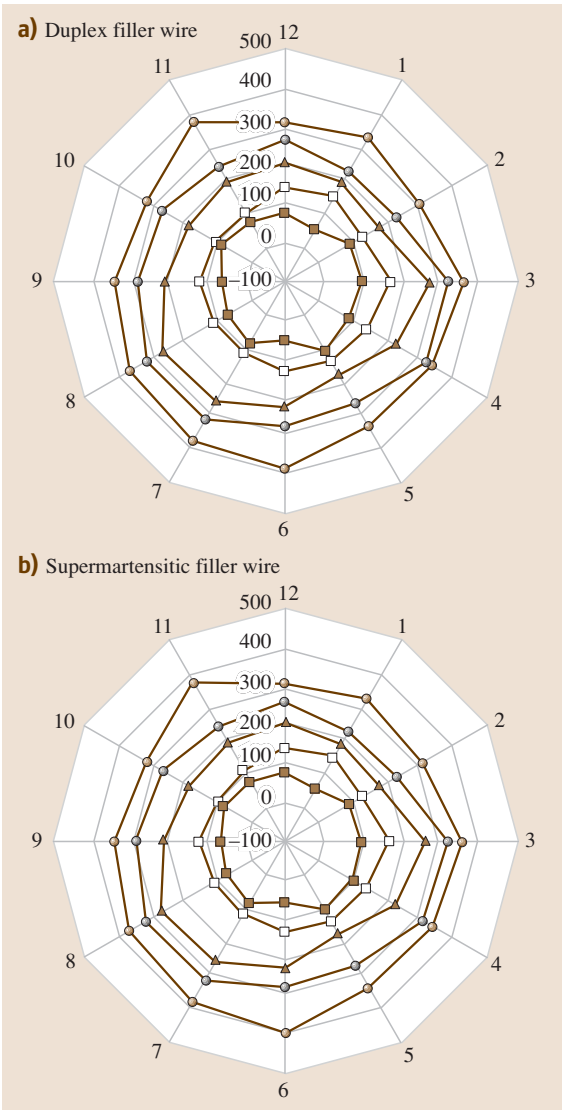


Fig. 7.292a,b Effect of selection of low temperature deforming supermartensitic stainless steel filler wire selection on the reduction of the circumferential stress distribution in an orbital pipeline weld (after [7.334]): **(a)** Duplex filler wire, **(b)** supermartensitic filler wire

sour aqueous environment, frequently containing hydrogen sulfide acting as a promotor.

With increasing application of high-strength structural materials and increasing mechanical loads introduced during fabrication welding, by lightweight design and respective high shrinkage restraints, but also during service by increasing acidity during prolonged life times of welded components, failure cases, some of catastrophic dimensions, originated by HACC or/and HASCC have been increasing recently. There is thus an increasing demand for modeling procedures to calculate hydrogen removal during welding pre- and postweld heat treatments, to assess crack propagation, and to predict the lifetime of welded components in corrosive environments.

Quite a number of analytical models have been developed to simulate HAC, predominantly HASCC, which are mainly based on stress–strain thresholds and, thus, have to be regarded as crack-no-crack models. Rates of HAC growth have less often, though still been modeled analytically [7.335]. A numerical procedure to simulate HAC has thus been developed [7.336, 337], considering the cracking process as the propagating failure of fictive tensile specimens alongside the crack path [7.338], each being represented by a finite element in the numerical procedure (Fig. 7.294a).

As shown by the flow chart in Fig. 7.294b, numerical modeling starts with an analysis of the hydrogen concentration profile to evaluate the local hydrogen concentration near the crack tip. In the second step, the local mechanical load in terms of local strains or stresses is analyzed. The critical hydrogen concentration in the crack initiation range or near the crack tip, respectively, is then calculated from the crack criteria, usually given as a database or as a mathematical relation of the material properties dependent on the hydrogen concentration. Crack initiation or propagation, respectively, appears if the local hydrogen concentration at the crack tip exceeds the critical hydrogen concentration. After a respective time and crack increment, the procedure starts from the beginning and repeats. As particular advantages, in contrast to analytical models, such a numerical procedure allows exact calculations of the crack increment per time interval and also considers correctly the geometric effects of crack propagation on the hydrogen redistribution at the crack tip. Such numerical procedures generally require exact input data extracted from experimental results, like the diffusion coefficient and, in particular, hydrogen-dependent material properties.

As shown by Fig. 7.295, such numerical modeling has meanwhile been successfully applied to HASCC [7.336] as well as to HACC [7.338] in welds of supermartensitic stainless steels used for offshore pipeline applications. During recent failure cases, it turned out that HASCC in critical regions sometimes was associated with previously undetected HACC phenomena. Thus, more recent investigations have focused on the evaluation of hydrogen diffusion and cracking

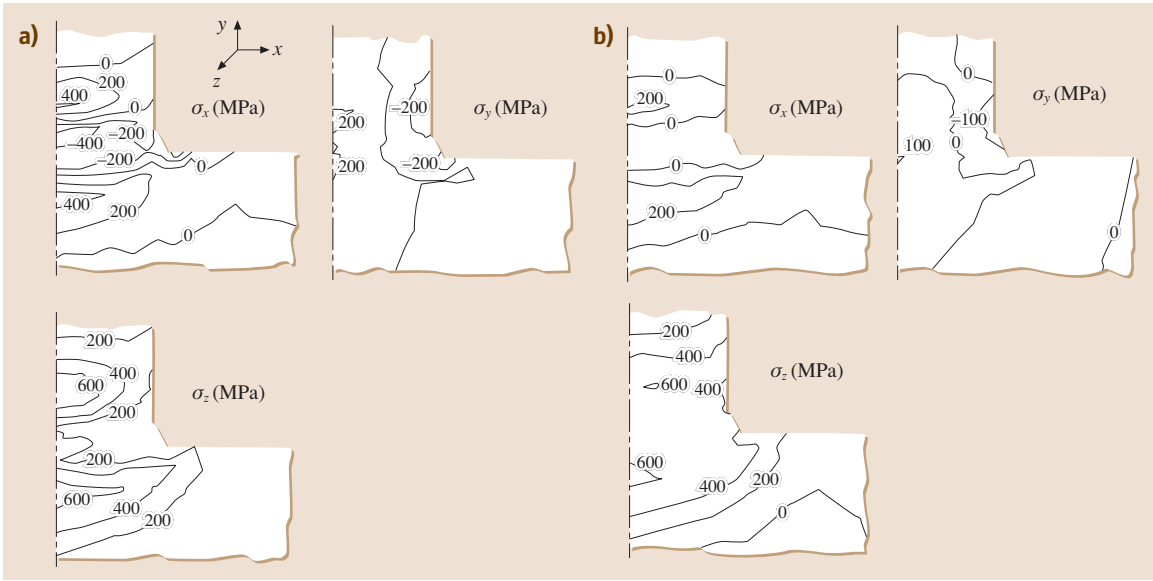


Fig. 7.293a,b Simulated stress fields in a half a symmetrical T-joint **(a)** without considering transformation plasticity and **(b)** with consideration of transformation superplasticity (after [7.332])

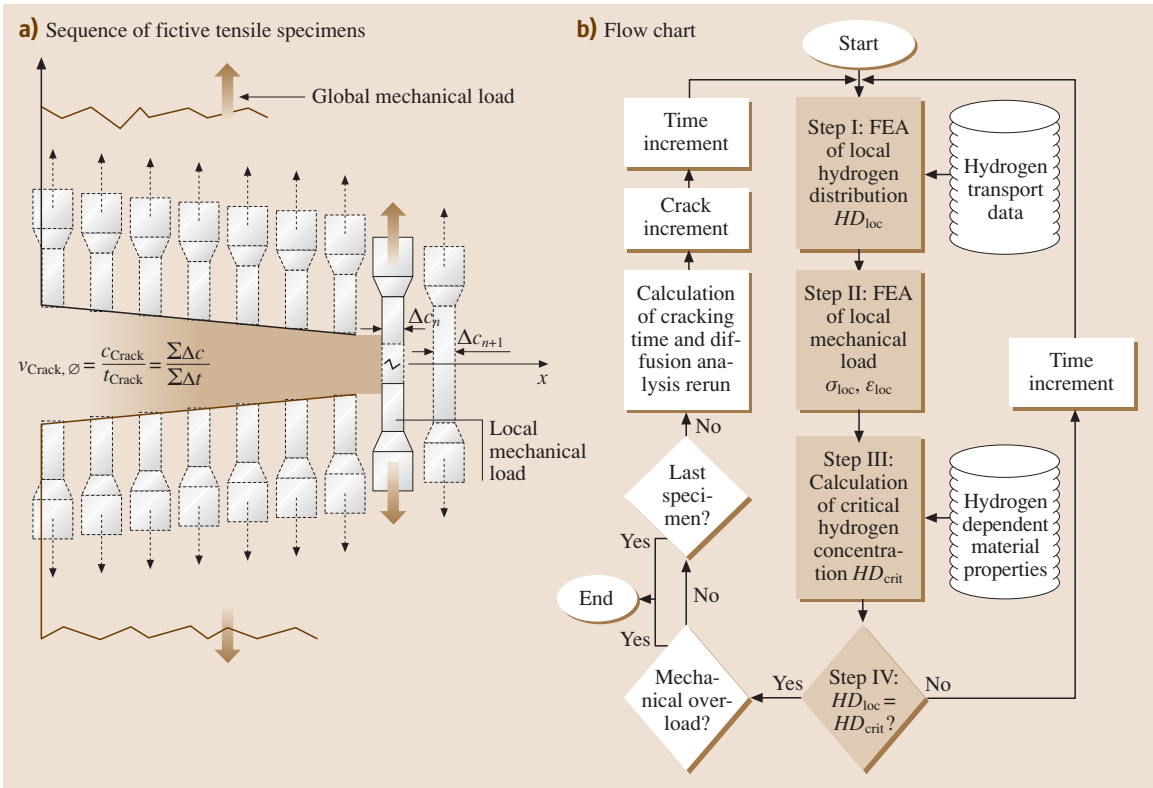


Fig. 7.294a,b Concept of numerical modeling of hydrogen assisted cracking (after [7.338]): **(a)** Sequence of fictive tensile specimens, **(b)** flow chart

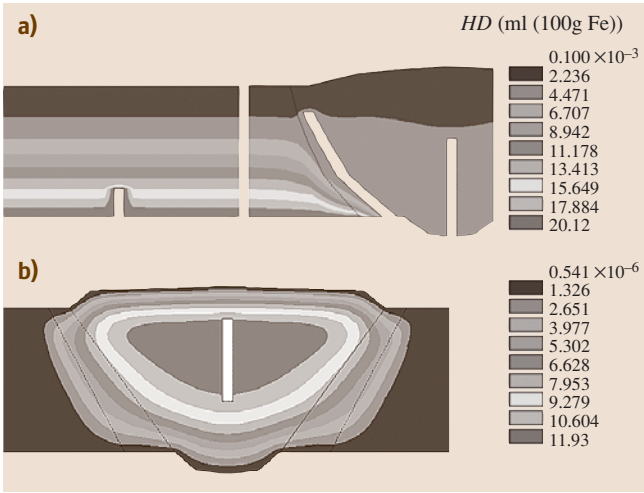


Fig. 7.295a,b Hydrogen Concentration profiles (ml/100 g) and crack propagation in supermartensitic stainless steel welds (after [7.338]): **(a)** Numerically modeled HASCC with predominant crack propagation in the heat affected zone, **(b)** numerically modeled HACC with predominant crack propagation in the weld metal

in high-strength structural steel welds [7.339], first providing the respective crack criteria and then developing numerical modeling aimed at crack avoidance and lifetime assessments of such components.

Liquid–Solid Transformations A fundamental understanding of weld solidification has evolved over time, based upon the application of general solidification models to the conditions specific to welding [7.340–342]. Included among these weld-specific conditions are epitaxial nucleation on base metal grains, continuous advancement of a liquid–solid interface, variable temperature gradient/growth rate (G/R) conditions along the liquid–solid interface, and high-temperature gradients (G). Each of these conditions has a marked effect on weld solidification structure and the development of defects:

- Epitaxial nucleation results in continuous columnar grains.
- Irregular advancement of the liquid–solid interface results in banding.
- Low G/R at the weld centerline results in grain refinement.
- High G/R results in fine dendrite spacing.

Due to the rapid cooling rate associated with welding, there is only limited time for diffusion of solute in the solid. Thus, nonequilibrium conditions apply to weld solidification, which typically occurs in either a dendritic or columnar-dendritic mode. Dendrite spacing has been shown to vary inversely with the square root of cooling rate [7.343, 344], where the cooling rate can represented by G/R. Thus, the scale of dendrites and microsegregation tends to be an order of magnitude smaller for welds than for castings (approx. 10 μm versus 100 μm), because cooling rates are much higher (approx. 100°/s versus 1°/s).

Models and simulations regarding solidification grain structure and microsegregation are discussed below. Also addressed are weld solidification defects including liquation and solidification cracking (both sometimes referred to as hot cracking) and porosity. Additional aspects of weld solidification phenomena can be found in general reviews on the subject [7.345, 346].

Grain Structure. Weld metal grains tend to be columnar and smaller in diameter than those found in castings (approx. 100 μm versus 1000 μm), in part because of the epitaxial nucleation of weld metal grains on small base metal grains. Grain growth competition also plays a role in this, wherein grains not having a favorable orientation (i. e., with dendrites parallel to heat flow direction) must grow at a higher undercooling. Such grains will eventually become occluded, either from adjacent grains or from nucleation of a new grain [7.341]. Thus, competitive grain growth tends to limit the growth of most columnar grains, although this appears to not always be case (e.g., austenitic stainless steels and titanium alloys exhibit exceptionally large weld metal grains).

Columnar grains grow normal to the weld pool boundary, as this minimizes energy associated with grain boundaries. Thus, the shape of the weld pool determines the growth path of the columnar grains [7.340]. The development of weld metal grain structure has been modeled and simulated using a variety of methods including cellular automata [7.347, 348], grain boundary evolution [7.349–351], and Monte Carlo [7.352], imposing boundary conditions for epitaxial nucleation, weld pool shape, and competitive growth. These simulations have proven useful in predicting weld metal grain structure for a given weld pool shape, as demonstrated in Fig. 7.296 [7.351].

Of particular interest to weldability, for improved mechanical properties and resistance to solidification cracking, is the ability to grain refine, i. e., replace columnar grains with equiaxed grains. Grain refinement requires a combination of high undercooling

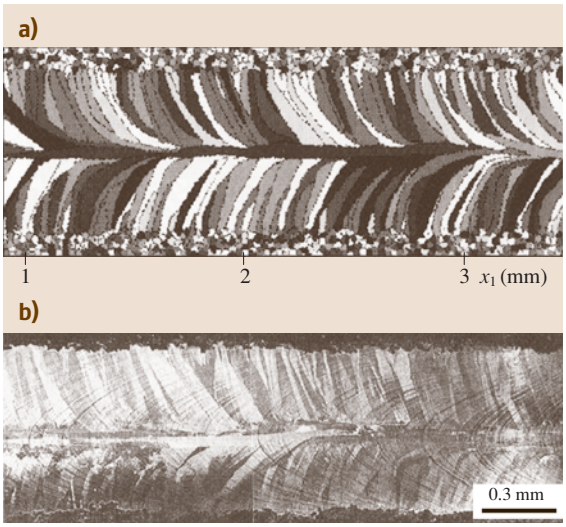


Fig. 7.296a,b Cellular automata simulation of weld grain structure **(a)** compared against actual weldment **(b)**, as viewed from top surface (after [7.351])

plus a substrate suitable for grain nucleation. Models used to predict the columnar-to-equiaxed transition (CET) must examine the undercooling ahead of dendrites in a columnar grain, which takes the following form [7.353]

$$\Delta T = \frac{GD}{R} + AR^{1/2}\,, \qquad (7.201)$$

where D is the solute diffusivity in the liquid and A is a material constant. Assuming site saturation for the nucleation of new grains ahead of the advancing columnar grains, equiaxed grains will nucleate at some critical undercooling, ΔT_N, and then grow until impingement by the columnar grains. CET can be achieved if these equiaxed grains have time to grow to sufficient size before impingement occurs [7.355]. This model has been applied to welding [7.356], where it has been shown that equiaxed grains can be expected at the weld center (high R) and along the fusion line (low R), i. e., locations of high undercooling.

Microsegregation. Modeling of microsegregation for nonequilibrium, dendritic solidification has been accomplished using the Scheil equation [7.357]

$$C_s = kC_0(1 - f_s)^{k-1}\,, \qquad (7.202)$$

where C_s is the solute concentration in the solid at the liquid–solid interface, k is the equilibrium partition coefficient, C_0 is the nominal solute concentration, and f_s is the solid fraction. When applied to a volume element oriented perpendicular to a cellular-dendrite axis (Fig. 7.297), this equation can be used to describe the distribution of solute across the core of the dendrite, assuming there is no diffusion in the solid.

In cases where corrections must be made for back-diffusion, a numerical finite difference solution becomes useful. An example of this is given in Fig. 7.298, showing the results of a numerical simulation using the Scheil equation, corrected for back-diffusion, for a Fe-21%Cr-14%Ni stainless steel alloy weld metal [7.354]. The nickel content is represented by various shades of gray and assuming solidification proceeds as either (a) primary austenite or (b) primary ferrite. In the case of primary ferrite, there is considerably more homogeneous distribution of nickel, due to a higher diffusivity.

Liquation Cracking. Liquation cracking, also known as microfissuring, occurs in the partially melted zone (PMZ) of the HAZ due to the separation of liquid films present in grain boundaries. Certain alloy systems are

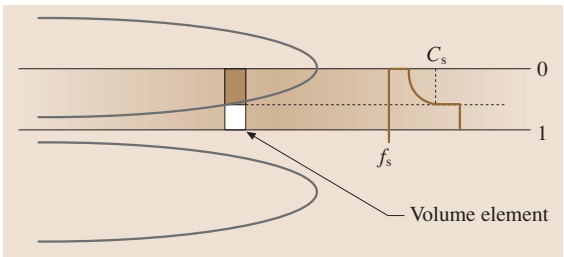


Fig. 7.297 Schematic of cellular-dendrite showing Scheil volume element and solute distribution within dendrite

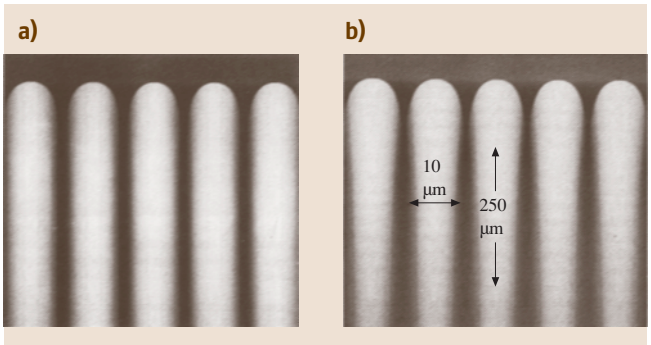


Fig. 7.298a,b Predicted distribution of nickel across cellular-dendrites in Fe-21%Cr-14%Ni solidifying as **(a)** austenite and **(b)** ferrite. Ni concentration increases with increasing darkness (after [7.354])

notably susceptible to this defect (e.g., nickel-based alloys and aluminum alloys). In nickel-based alloys (e.g., alloy 718), problems with liquation have been tied to the presence of niobium carbide (NbC) particles [7.358]. Aluminum alloys of the type Al-Mg-Si are known to be particularly problematic with regard to liquation [7.359, 360]. Even when there is insufficient strain to cause cracking, microstructural damage caused by liquation alone can impair joint performance.

One early model for liquation focused on nonequilibrium conditions, relating melting to the partial dissolution of residual eutectic particles present at grain boundaries, better known as constitutional liquation [7.362]. As solute diffuses from grain boundary particles into the matrix, melting will occur at a point where the eutectic composition is reached, provided the eutectic temperature has been exceeded. This assumes that the particle does not completely dissolve before the eutectic temperature is reached. The rapid heating and cooling rates associated with welding are particularly conducive to this mechanism as demonstrated in a coupled heat and mass flow model [7.363]. The liquid film produced during constitutional liquation can spread along grain boundaries and promote constitutional liquid film migration (CLFM) as observed in the nickel-based alloy 718 [7.358].

In more recent studies on aluminum-copper alloys [7.364], it has clearly been shown that equilibrium liquation is also possible whenever the solidus temperature is exceeded, although the kinetics of such reactions are not well established. It was demonstrated that cracking is likely to occur only under specific solidification conditions, where the weld metal fraction solid is consistently above the partially melted zone fraction solid [7.365]. This suggests that a strong weld metal, capable of resisting strain, will transfer strain to the HAZ where cracking can then initiate. Curves representing solid fraction were calculated based upon the Scheil equation (7.202).

Solidification Cracking. Solidification cracking involves the separation of liquid films present at grain boundaries in the mushy (liquid + solid) zone trailing the weld pool. Both thermal strains and solidification shrinkage act in conjunction to pull grains apart, initiating and sustaining crack growth. Certain alloys (e.g., fully austenitic stainless steels, steels having high sulfur or phosphorus impurity levels, and high-strength aluminum alloys) are known to be particularly susceptible to this type of cracking. A large solidification range and a continuous, columnar grain structure appear to favor cracking [7.366].

Numerous tests have been developed to asses cracking susceptibility, including Varestraint [7.367], circular patch [7.368], sigmajig [7.369], and PVR [7.370] tests. These tests rely upon large, reproducible, externally augmented strains (extrinsic type test) or internally generated strains (intrinsic type test) to generate extensive cracking for comparative purposes. These tests allow weldability to be quantified in terms of crack length and can thus be used to compare the cracking susceptibility of different base metals, filler alloys, and welding procedures. Cracking is often quantified in terms of total accumulated crack length (TCL), although the exact meaning and relevance of this value is questionable.

Most cracking models have centered on defining a brittle temperature range, and the maximum tolerable strain within this range, for predicting the behavior of a given alloy [7.371, 372]. At its extreme, this temperature range is bounded by the liquidus and solidus, but it has also been taken to be more restrictive, e.g., a coherent temperature range, whereby dendrites have interconnected and can resist strain. More recently, the maximum length of crack generated in a Varestraint test has been used to identify a critical temperature range for cracking, i. e., solidification cracking temperature range (SCTR), which may reflect more on an inherent propensity for cracking than TCL [7.373].

Other approaches have examined the conditions necessary for crack initiation based upon a drop in in-

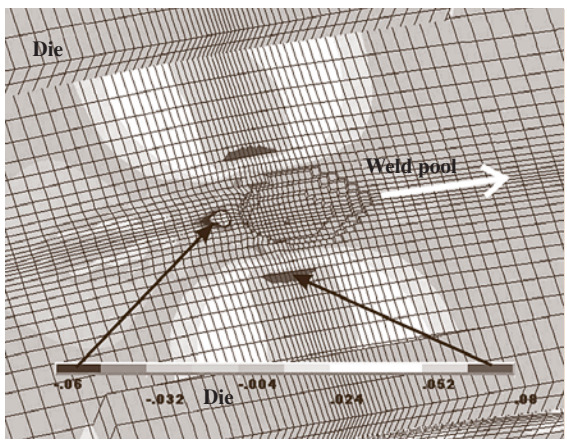


Fig. 7.299 Computer simulated strain distribution in a modified Varestraint test (strain in direction of welding) showing complex distribution of tensile and compressive strains beside and behind a moving weld pool after the test strain has been applied (after [7.361])

terdendritic liquid pressure, relating this to an inability to adequately feed solidification shrinkage and thermal contraction [7.374–376]. Although these models were developed for castings, they are equally applicable in principle to welding. Of particular importance to welding, however, is the influence of restraining conditions and their effect on local strain fields around a moving weld pool [7.377]. The ability to simulate these local stress/strain fields is of critical importance for predicting cracking, with several examples published [7.361, 378–380]. An example of one such simulation is given in Fig. 7.299, showing the strain distribution in a modified Varestraint test [7.361].

Porosity. Gas porosity in metals is usually associated with dissolved interstitial elements: H, N, and O, which form gas (diatomic molecules) during solidification. The conditions necessary for gas pore formation in molten metal have been modeled based upon a simple consideration of Gibbs free energy for nucleation [7.381]

$$\Delta G = \gamma A + P_e V - P_i V \,, \tag{7.203}$$

where ΔG is the change in free energy associated with pore formation, γA is the energy associated with the creation of gas/liquid interface of surface area A and surface tension γ, $P_e V$ is the positive work done forming a pore of volume V against an external pressure P_e, and $-P_i V$ is the negative work done with the aid of internal vapor pressure P_i. From this it follows that the critical radius needed for homogeneous nucleation of a spherical pore (i. e., r such that $\Delta G/\mathrm{d}r = 0$) takes the following form

$$r^* = -2\gamma/\Delta P \,, \tag{7.204}$$

where $\Delta P = (P_e - P_i)$. It follows that homogeneous nucleation is promoted by low surface tension, low external pressure, and high internal pressure (e.g., EB welding a nitrogen strengthened stainless steel in a vacuum). However, heterogeneous nucleation is even more likely to occur, allowing this critical radius to be achieved with a smaller volume of gas. Heterogeneous nucleation is favored by conditions where the liquid does not wet the substrate (e.g., oxide inclusions [7.382]).

ΔP can be considered as the driving force for nucleation and can thus be expressed in terms of a thermodynamic chemical potential [7.382]

$$\Delta P = -\frac{\mathrm{RT}}{\Omega} \ln \frac{p}{p_0} \,, \tag{7.205}$$

where R is the universal gas constant, T is absolute temperature, Ω is the molar volume, p is the interstitial partial pressure, and p_0 is the equilibrium interstitial partial pressure. It follows that there must be a condition of supersaturation (i. e., $p > p_0$) in order for nucleation to occur. Thus, in order to understand pore formation in a weld, one must first consider mechanisms to achieve supersaturation.

Pores can form in the weld pool if the liquid becomes supersaturated by picking up interstitials from the welding gas, directly under the arc (hot region), and then moving rapidly to cooler regions near the fusion line. The pickup of interstitials at the weld pool surface has been modeled, where it has been shown that dissolved interstitial concentrations exceed Sieverts' law predictions due to the dissociation of diatomic molecules in the arc plasma [7.383]. It is also possible to form porosity interdendritically, even when the weld pool has not become saturated, due to the partitioning of interstitial elements during solidification. The large drop in solubility between liquid and solid results in a buildup of interstitials in the interdendritic liquid, which leads to supersaturation. Simulations for interdendritic porosity have been used successfully to predict conditions favorable to pore formation in castings [7.384, 385].

Once nucleated, pores may grow, coalesce, become entrapped, or escape the weld pool, depending upon the welding conditions [7.382]. Slow welding speed allows time for pores to escape, aided by the buoyant force of gravity. The use of pool agitation (e.g., current pulsation) also helps in this regard. When welding in the overhead position, most pores become entrapped. Rapid travel speed limits the time available for pore nucleation and growth.

7.4.9 Fundamentals of Magnetic Pulse Welding for the Fabrication of Dissimilar Material Structures

One major challenge in welding is to develop fast, reliable, and cost-effective industrial processes to permanently join dissimilar materials, i. e., different metals (including alloys), or metals with plastics or ceramics. For these combinations of materials, the fusion welding processes are inapplicable, as the physicochemical properties of unlike materials are seldom similar or *compatible*. Alternatively, the *colder* and solid-state joining processes like magnetic pulse welding (MPW) offer the most potential, particularly for cylindrically symmetrical components including light and ductile

metals. MPW is straightforward and in many aspects similar to explosive bonding, except that it has for now only been applied to cylindrically symmetrical components. MPW utilizes the magnetic fields generated by heavy discharge currents into inductive coils. The resulting discharge is a dampened sine wave of consecutive *pulses*. In proximity to the coils are the components to be welded, or workpieces. The discharge current running through the coils induces Eddy currents in the nearby workpiece. The interactions between the magnetic fields of these two currents results in a strong repel force between coil and workpiece. By necessity, this workpiece must be electrically conductive as well as plastically deformable, and the repel forces must be such that a violent collision will occur, preferentially at a slight angle to form a jetting action similar to that in explosive welding. During MPW, the amount of heat produced is almost nonexistent since the process only lasts small fractions of a second. Parameters such as gap distance (between coil and workpiece, and between the two workpieces), material properties, thickness, as well as welder characteristics determine the properties of the final weld joint. Mechanical interlocking as well as thin and discontinuous intermetallic phases (all resulting from localized melting) will control the final properties of dissimilar-material welds, e.g., static strength, shock and vibration resistance, and vacuum tightness.

Magnetic pulse welding (MPW), also referred to as electromagnetic impulse joining or pulsed magnetic welding [7.386], is a four-decade-old process from the Cold War era [7.387]. Just like other governmental programs, the widely developed nuclear programs of the former Soviet Union, the United States, and other military and industrial powers spawned spinoff technologies that have found industrial and manufacturing applications. MPW is said to have been invented at the Kurchatov Institute of Nuclear Physics to seal metal canisters and nuclear fuel pins [7.387]. In the decade following its initial success until recently, the process only found military and aerospace applications, as for flight control rods, artillery shell casings, and bimetallic metal inserts [7.388, 389]. Almost 40 years after being invented, MPW has gained the attention of the private sector, in particular the transportation and refrigeration industries. With weight saving and improved vehicle safety driving the use of an increasing number of dissimilar materials' joints (e.g., aluminum with steel), the automotive community has emerged as the major player in further developing MPW technology. MPW is indeed one of the rare processes capable of joining dissimilar materials in high-volume production environment. The actual process lasts less than 100 ms, and the production rates may be readily customized (for instance, be as small as a few seconds).The process has not only been tested and applied to numerous combinations of metals and alloys [7.388–398], but also metals with ceramics or metal-matrix composites [7.399]. In the automotive industry, immediate potential applications for MPW include more conventional metallic materials for air conditioning tubings, tubular spaceframes, driveshafts, struts, shocks, and electrical connections. To date, any joints between round parts such as a tube-to-tube joint, a tube-to-end joint or a wire crimp joint are ideal candidates for magnetic pulse welding. In the near-future, it is conceivable to see magnetic pulse welding, alone or combined with other processes, applied to noncylindrical workpieces like flat sheets. MPW is an extension of magnetic pulse forming, or electromagnetic forming, a process that uses identical technology to manufacture complex shapes in fractions of a second.

Process Principles and Parameters

MPW can be applied to the same materials as explosive bonding, provided the hollow sections can be accelerated. MPW is identical to explosive bonding in the formation of the joint, but instead of the chemical explosive energy, magnetic fields are used to drive the materials together. In order to weld and in particular achieve a metallurgical bond wherein atoms of the two materials are brought into direct contact, a tremendous amount of energy must be compressed and discharged within an extremely short time. In some systems, the discharge is as high as 2 million amps and lasts less than 100 µs. As a result, the actual energy expenditure is exceptionally low and the components have no time to heat appreciably.

A schematic representation for a magnetic pulse welder is illustrated in Fig. 7.300. The welding unit, or welder, simply consists of an LC circuit (i. e., inductance–capacitance) with a high-voltage transformer and some impedance (not represented in Fig. 7.300) so that the discharge current waveform is a dampened sine wave. This discharge current runs through a coil, also called an inductor. This coil is usually positioned all around the workpieces to be welded, but not necessary as discussed later in this chapter. Figure 7.300 illustrates the situation where two concentric workpieces (i. e., an internal and an external workpiece) are welded. The workpieces are positioned coaxially inside the coil with gaps in between, a requirement to produce the Eddy currents and the magnetic force necessary for welding. The electrical currents in the coil

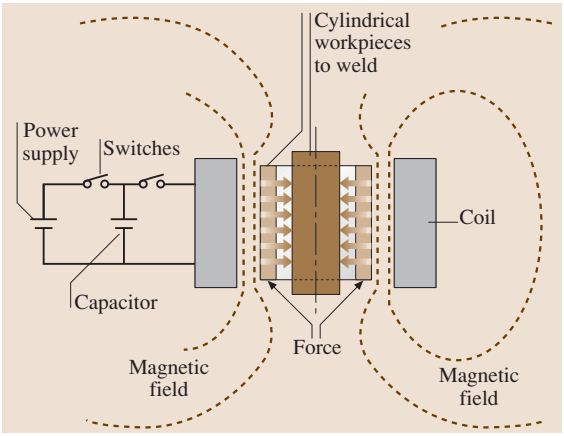


Fig. 7.300 Basic components and principle behind the magnetic pulse welding process

and workpiece (i. e., the inductive and the Eddy currents) naturally repel one another since the two currents, and their respective magnetic fields, are – as controlled by the laws of electromagnetism – flowing in opposite directions. This feature makes MPW comparable to what happens when two permanent magnets are held together, i. e., when like poles are brought to close distances they repel each other. Instead of expanding and penetrating each other, the two opposing magnetic fields develop a pressure that puts one of the workpieces into sudden motion. Once the distance separating the two workpieces, or gap, has been covered, the moving workpiece finally collides with the stationary workpiece (i. e., the inner workpiece in Fig. 7.300). The average velocity is typically on the order of 300 m/s [7.392, 393], but the local velocity can be significantly higher (e.g., greater than 1000 m/s) [7.388]. Later, we will discover that conversion of this kinetic energy to plastic deformation often results in local heating; as a result, magnetic pulse welds may develop a variety of characteristics depending upon the materials and established conditions of collision.

To create conditions for welding, two conflicting machine characteristics are required: a large energy storage, which usually means a high capacitance (produced by a bank of capacitors), and a high frequency, which means a low inductance. The energy storage E and the frequency of an LC circuit are given by [7.386]

$$E = \frac{1}{2} C U^2 \,, \tag{7.206}$$

$$f = \frac{1}{2\pi\sqrt{LC}} \,, \tag{7.207}$$

where C is the capacitance (in F), U is the charging voltage (in V), and L is the total inductance of the unit (in H). Since frequencies for MPW are usually on the order of 10–200 kHz, the magnetic field produced by the inductive coil only diffuses through the skin of the nearer workpiece, a phenomenon sometimes referred to as the Kelvin effect [7.400]. The higher the resistivity of the material and the lower the frequency, the more penetrating the induced current and the less effective the induced magnetic field. For the situation described in Fig. 7.300, the induced current will result in a high magnetic field on the outer surface of the exterior workpiece and little or no magnetic field on its interior surface, an imbalance that will cause the collapse of the external workpiece. The velocity when this workpiece impacts the stationary workpiece depends on the magnetic pressure, the mass of the accelerated workpiece, its material properties, and the initial gap between the two workpieces. The magnetic pressure, created by the inductive current pulses or oscillations, can be approximated as [7.398, 401]

$$P = \frac{\mu_0 K^2 n^2 U^2 C \sin^2(\omega t) \exp(-Rt/L)}{2 L I_\mathrm{w}^2} \,, \tag{7.208}$$

where μ_0 is the magnetic permeability (in H/m), K is a coefficient depending upon the physical dimensions of the coil, n is the number of windings of the coil, ω is the LC oscillation frequency associated to (7.207), R is the total resistance of the discharge circuit (in Ω), I is the working coil length (in m), and t is the time (in seconds). Because current and pressure rapidly dampen, the collision between the two workpieces must take place in about half the period of current oscillation to maximize the process efficiency, in particular the collision velocity.

To illustrate the effects of the various process parameters, a simplified description of the process can be made by combining several equations, starting with Newton's second law, which is applied to describe workpiece motion as a function of the time and the various forces developing during the process

$$m \frac{\mathrm{d}^2 h}{\mathrm{d}t} = F_\mathrm{m} - F_\sigma \,. \tag{7.209}$$

In (7.209), m is the mass of the moving workpiece (in kg), h is the gap separating the two workpieces (in m), F_m is the induced magnetic force (in N), which can be approximated by multiplying magnetic pressure (7.208) by external area of the outer workpiece (for the configuration illustrated in Fig. 7.300), and F_σ is the force

induced by the accelerated workpiece as it resists deformation (in N). This force may be described as [7.398]

$$F_\sigma = \sigma 2n l_w r \,, \tag{7.210}$$

where σ is the plastic flow stress of the moving workpiece material (i. e., its strength in N/m^2 or Pa), l is the length of its deformed area (in m), and r is either the outer radius or inner radius (in m) of the moving workpiece (whichever applies best). In (7.210), σ not only represents the strength of the moving material under uniaxial conditions (a simplification), but it is also taken independently of the developing microstructures, i. e., strain hardening and strain-rate dependence must be neglected. Equations (7.208–7.210) infer that the magnetic force must first overcome the resistance of material and force it to plastically deform. Although not represented by any equation here, the magnetic force must also remain below a critical value, above which too much of the work produced by the magnetic pressure will result in heat generation during the collision, a situation that would eventually cause interfacial melting and cracking, as discussed in several publications [7.394, 395, 397] and later in this chapter. Equations (7.206) and (7.208) indicate that the charging voltage must also be high to increase the amount of stored energy in the capacitance, and thus produce the adequate magnetic pressures for welding. Typically, given the fixed design of a magnetic pulse welder, a major factor influencing the quality of a weld assembly is the geometry or design of the weld joint. Equation (7.209) suggests that the gap between the coil and the workpiece and between the workpieces (represented by h in (7.209)) is a key parameter. Wider gaps will provide the time to accelerate the moving workpiece and thus create the high collision velocities that are needed for welding.

Process Requirements

MPW requires a lap joint configuration with several important characteristics in order for high-quality joints to be achieved. Herein, we will define high-quality joints as those that satisfy a set of desired properties. While mechanical strength (static, dynamic, i. e., under impact conditions, or fatigue) is always important for product reliability and durability, other properties may be equally important in certain applications. For instance, when excellent electrical and thermal conductors like aluminum and copper are bonded, conductance may be primary, and the detailed characteristics of the weld (described later in this chapter) may then become important to consider. For MPW, the workpieces may be of a variety of sizes. They may be as small as a few millimeters in diameter, but not excessively small to prevent the magnetic fields from interacting with each other. Workpieces can be as large as the application requires and the magnetic pulse welder allows. Workpieces in excess of 100 mm are not yet common but are weldable by magnetic pulse technology with the appropriate equipment and joint design.

As discussed earlier, the gap that initially separates the moving workpiece from the stationary workpiece directly affects the collision velocity (in (7.209) dh/dt represents the velocity). This gap can be either constant or variable, as in the cases of tapered workpieces. Figure 7.301 depicts a variety of magnetic pulse weld joints. The configurations with tapers offer considerable advantages, especially in producing the jetting described in the next section. Figure 7.301 also reveals that the coils can be placed either on the inside or on the outside of the workpieces. With external coils, the moving workpiece collapses onto the inside stationary workpiece. For an inside or internal coil, the inner workpiece will expand radially and collide with the inside of the wider and peripheral workpiece. The choice of coils, internal or external, depends on the dimensions of the assembly, as well as on the materials and weld properties that must be achieved. External coils can be provided to weld outside diameters between ≈ 2 and 100 mm. Internal coils are typically used for diameters between 100 and 250 mm.

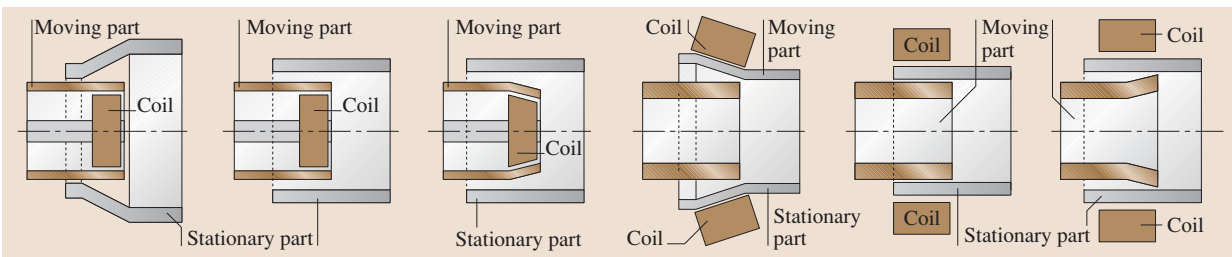


Fig. 7.301 Magnetic pulse weld configurations with corresponding positioning for the inductive coil (after [7.398])

The properties of the workpieces will affect the design of the joint. For instance, the workpiece that is to be accelerated must be made of a good electrical conductor. The higher its electrical conductivity, the easier it can be moved provided its section is relatively thin and deformable at high strain rates. Ideally, the more conductive material is also the least strong and most ductile. Metallic materials that have an electrical resistivity of 15 μΩ/cm or less lend themselves well to magnetic pulse welding. Some of these materials include alloys of aluminum, copper, and the precious metals. When welding electrically resistive materials, the more conductive of the two should preferentially be placed near the coil and this coil may be located on the outside to yield the necessary magnetic forces. This situation describes, for instance, the welding of an aluminum tube over a copper, titanium, or steel workpiece. In Fig. 7.302 are depicted several such weld joints before and after peel testing (peeling is used as a simple destructive test to assess joint strength) [7.395]. Aluminum-to-copper and aluminum-to-stainless-steel joints involving tubes of comparable wall thickness are shown. Because stainless steel's plastic flow stresses are significantly greater than that of aluminum and its alloys, the inner tube does not collapse on itself during the collision. With steel on the inside, the final component is undistorted and generally within the dimensional tolerance imposed by the design. Characteristics of several of the joints of Fig. 7.302 will be revisited later in this chapter. As can be seen in Fig. 7.301, the workpiece in close vicinity to the coil will often be forced to stretch before contacting the stationary workpiece. As a result, MPW inherently involves a significant amount of plastic deformation prior to bonding. While many metals and alloys are highly deformable, especially in the high strain rate and adiabatic conditions created by the process, it is clear that hard and very brittle materials may be unsuitable (e.g., tungsten).

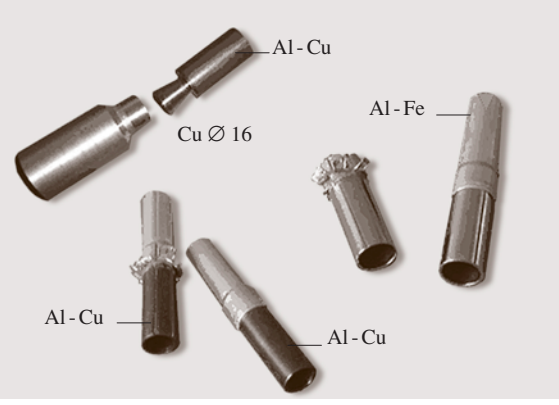


Fig. 7.302 Examples of hybrid structures of aluminum fabricated by magnetic pulse welding

In developing MPW technology, the electrical conductivity of metals and alloys has imposed severe requirements on the capacitor bank, the load inductor, and the tooling [7.402]. For welding, electrically resistive metals such as stainless steel, titanium, or nickel alloys necessitate a high ringing frequency (and consequently a low inductance is required). Less electrically resistive metals like aluminum and copper can be welded satisfactorily with lower frequency and lower current machines. Of primary importance, the coils for a magnetic pulse welder must be significantly stronger than the workpiece to be projected in order for them to be reusable. Many of components must also be electrically isolated so that the fields generated by the coil do not interact with any adjacent components. Because of the close tolerances required between the coil and the workpieces, and in between the workpieces, coils are frequently designed unique to their applications, i. e., tube size, rigidity, and materials.

While a thinner moving workpiece is preferred for a rapid acceleration before collision, a thicker stationary workpiece is often necessary to avoid distortions on the inner workpiece and keep the entire assembly within the imposed geometrical tolerance. MPW of thin-walled tubes on other thin-walled tubes is therefore difficult, unless a mandrel is used to back up the inner workpiece. Since magnetic pulse welds are lap welds, the length of the overlap is a critical parameter to achieve high strengths, particularly high tensile-shear strengths. Difficult-to-bond materials will require longer overlaps than easier-to-bond materials. Also, for the long welds sometimes needed by products, split coils may be designed to allow easy loading and removal of long components.

Characteristics of Magnetic Pulse Welds

Jetting. Welding may be attributed to two major mechanisms. In order to produce strong welds, the two workpieces must be brought close to each other at the atomic level, i. e., the two materials must at least be allowed to diffuse into each other [7.392–398]. In a normal atmospheric environment, in which MPW is usually conducted, the metallic surfaces are covered with thin films of oxides, nitrides, and other adsorbed gaseous elements. For instance, aluminum is covered with alumina and copper with a variety of oxides with

compositions and crystallographic structures that depend upon prior thermal exposures. These surface films are diffusion barriers and will prevent materials from the two workpieces from exchanging atoms and thus create a true metallurgical bond. Fortunately, with a few exceptions, these films do not have to be removed before welding because of the phenomenon known as jetting [7.403]. Jetting occurs when the two workpiece collide. As one strikes the other, a high pressure is generated ahead of the collision and the metal in the vicinity is largely swept away, giving rise to this scouring jet of semimolten materials. As illustrated in Fig. 7.303, the jet contains fragments of the surfaces from both workpieces and leaves a virgin metal in close contact under very high pressure. As also depicted in Fig. 7.303, the materials also become severely plastically deformed along the weld interface. The interface of a magnetic pulse weld will often contain large numbers of projections and depressions with more or less periodicity, depending upon prior surface preparation and variations in gap distance [7.404]. These irregularities contribute to the high strength (i. e., fracture forces) of magnetic pulse welds. They not only increase the effective area of the joint, but also interlock the two surfaces, thus providing another bonding mechanism that will contribute to the high strength of many magnetic pulse welds.

The major parameters that determine whether a high-quality joint, preceded by the jetting phenomenon, will occur are the collision angle and the collision velocity, which are both dependent upon the magnetic field strength, the workpiece thickness, and its characteristics. With materials of relatively low plastic flow stresses and high ductility, the collision angle does not need to be as high as with strong and less ductile materials. Equation (7.211) relates collision angle with a material's hardness and collision velocity. Although (7.211) was initially developed for explosive bonding

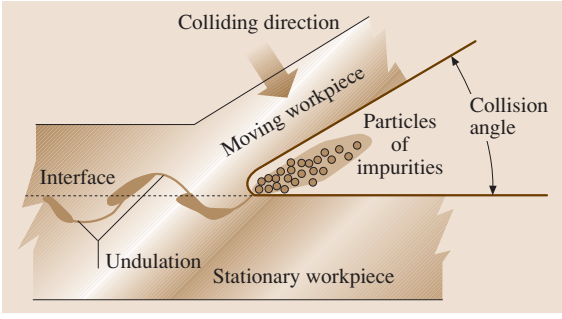


Fig. 7.303 Schematic representation of the jetting process for two workpieces entered in contact

to select the appropriate collision angle for jetting and bonding conditions, the same equation can be applied to MPW. Equation (7.211) indicates that the minimum angle for jetting to occur increases with the hardness of the moving workpiece H_v and decreases as its density ρ and its collision velocity V_c decrease [7.404]

$$\beta_{\min} = K\left(\frac{H_v}{\rho V_c^2}\right) . \qquad (7.211)$$

In order to improve jetting and bonding, many workpieces are machined to provide such a minimum collision angle, as shown in Figs. 7.302 and 7.303. Typical collision angles are in the vicinity of 8–12°. For difficult-to-weld combinations of materials like those between transition metals, surface preparation prior to cleaning may be recommended in order to lower the velocity for jetting. Surface roughening by mechanical or chemical means (i. e., etching) may be employed, as are a few commercial chemical agents designed for explosive bonding.

Weld Characteristics. Up to this point, we have emphasized that MPW applies best to dissimilar metals and alloys that (1) are good electrical conductors and (2) exhibit measurable differences in plastic flow stress. Light metals (e.g., aluminum, magnesium, etc.) are therefore most appropriate for welding with transition metals (e.g., steel, titanium, etc.) despite oxides and the fact that brittle intermetallic phases form when these metals are in contact at elevated temperatures. As discussed earlier, the quality of the weld depends on process parameters and material properties and can be evaluated through various techniques, detailed later in this chapter. Of all these techniques, visual examination of a weld interface from cross-sections is particularly useful for determining whether a true metallurgical bond has formed. When examining dissimilar material welds from polished cross-sections, the interface between the workpieces can take one of three appearances: straight with direct metal-to-metal contact, straight with a continuous layer between the workpieces, or wavy [7.395–404]. These three morphologies are often present in the same weld joint, especially when one of the workpieces is tapered and a gradient of collision velocities is consequently established.

Figure 7.304 presents a set of optical micrographs along the interface of a magnetic pulse weld between aluminum and titanium. From left to right, the gap that initially separated the aluminum and titanium workpieces gradually increased and therefore created greater and greater collision velocities. As seen in Fig. 7.304,

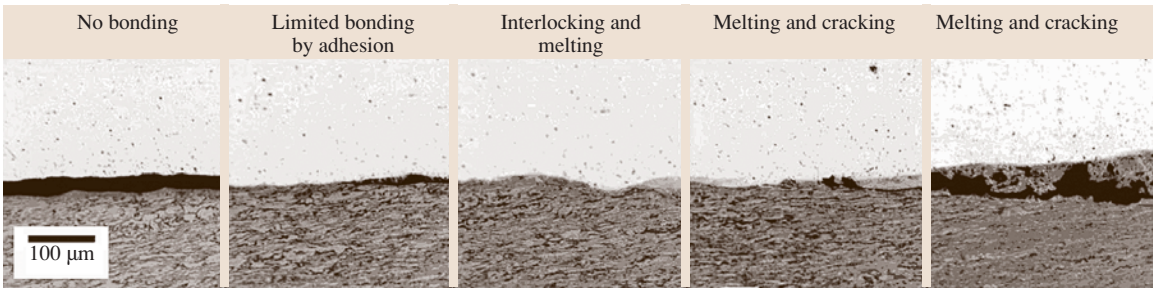


Fig. 7.304 Set of optical micrographs showing various interfaces at aluminum-titanium magnetic pulse welds. Collision velocity increased from left to right (after [7.395])

the narrower gaps of the left images produced no or only a limited bonding (attributed to adhesion), as can be seen in the first two images. As the distance between the workpieces increased, the two materials became slightly interlocked. To increase mechanical interlocking, larger, more penetrating, and more frequent surface irregularities would be necessary, as promoted for instance with surface roughening prior to welding. As the gap further widened, an interfacial microconstituent formed, exactly as seen in other dissimilar metal welds (e.g., aluminum with copper or steel) [7.388–390, 393, 394]. With the widest gaps, the interfacial microconstituent thickened, and cracking eventually developed all along the observed weld interface, as seen on the right-hand images of Fig. 7.304. Among the micrographs of Fig. 7.304, the conditions associated to the third image are best for the fabrication of high-quality aluminum-titanium welds. The ideal weld would be one with a high density of undulations for interlocking and with the thinnest interfacial microconstituent. This condition, although established by a narrow process window, has been observed to generally promote the highest mechanical strength and, if electrical and thermal conductivity are of concern, also the least resistance for current or heat to flow. The presence of an interfacial microconstituent is not preferred, but it guarantees that the two materials have truly created a metallurgical bond.

The presence of a new microconstituent, as illustrated in Fig. 7.304, is characteristic of many welds between dissimilar metals and alloys [7.387–398]. Figure 7.305 shows compositional maps captured at the cracked region of several dissimilar-metal welds. In the first, second, and third rows are backscattered electron images and compositional maps for an aluminum-low-carbon-steel weld, an aluminum-stainless-steel weld, and an aluminum-titanium weld, respectively. The aluminum maps are shown in the second column, and in the third column are the maps for iron and titanium. Using a particular colorscale, each of these maps represents element concentrations over a predetermined area, the brightest colors depicting the regions with highest concentrations. It is seen in Fig. 7.305 that the presence of a thick microconstituent is also associated with cracking. All the images of Fig. 7.305 also demonstrate that the various interfacial microconstituents are composed of at least one new intermetallic phase, perhaps more.

Line-scan results across the interfacial microconstituent of Fig. 7.304 (also represented in the third row of Fig. 7.305) are shown in Fig. 7.306. The first and last 40 µm on Fig. 7.306 indicate the chemical compositions of the two base materials, both expressed in

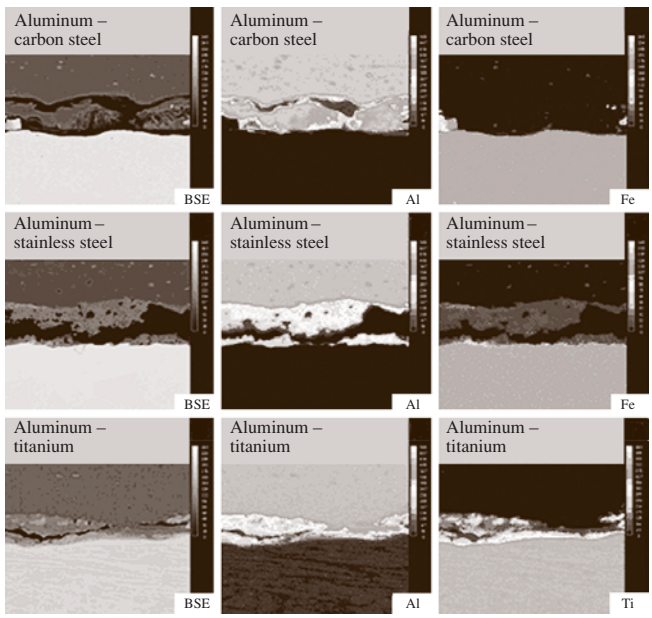


Fig. 7.305 Set of compositional maps reveling new phases at magnetic pulse welds between dissimilar materials

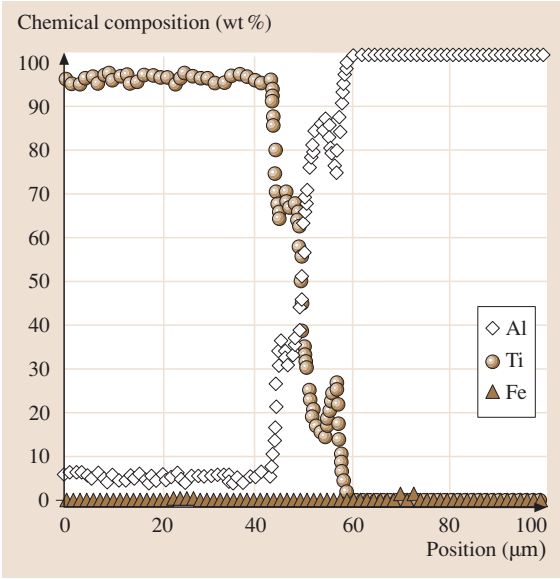


Fig. 7.306 line scan results showing the chemical composition across the interfacial microconstituent seen in aluminum-titanium magnetic pulse welds

weight percent. The chemical compositions of the interfacial microconstituent are found at distances between 40 and 60 μm on the *x*-axis. In this region, steep compositional gradients prevailed, except near 45 and 55 μm, where the compositional gradients changed sign and zeroed, thereby indicating the absence of a chemical driving for diffusion. In two narrow zones, aluminum and titanium have reached a local equilibrium and have therefore formed new phases. Figure 7.306 reveals that compositions of these two new phases were approximately 62–70 wt % titanium (i. e., 30–38 wt % aluminum) for the phase seen at a distance of 45 μm, and 15–25 wt % titanium (75–85 wt % aluminum) for the phase at a distance of 55 μm. It is here confirmed that the chemical elements of different workpieces do not participate equally in formation of the joint. It is also shown that the two new phase compositions match two intermetallic phases on the equilibrium phase diagram between aluminum and titanium [7.403]. This phase diagram is reproduced in Fig. 7.307. It shows that up to four equilibrium intermetallic phases can coexist between aluminum and titanium at temperatures less than 500 °C. They are $TiAl_3$, $TiAl_2$, TiAl, and Ti_3Al, several of which can be stabilized over a significant range of composition. As a consequence of four intermetallic phases, five ranges of compositions are available to produce dual-phase microstructures, i. e., microstructures with two phases (in Fig. 7.306, these dual-phase microstructures were found in a region of compositional gradients). The binary-phase diagram of Fig. 7.307 for titanium and aluminum resembles many other binary-phase diagrams between aluminum and a transition metal element, like iron (steel) for instance. As a result of such similarities in the phase equilibria, interfacial microstructures in aluminum-carbon steel or aluminum-stainless steel welds are not profoundly different from aluminum-titanium welds, and thin layers of intermetallic phases are produced, as demonstrated in Fig. 7.305.

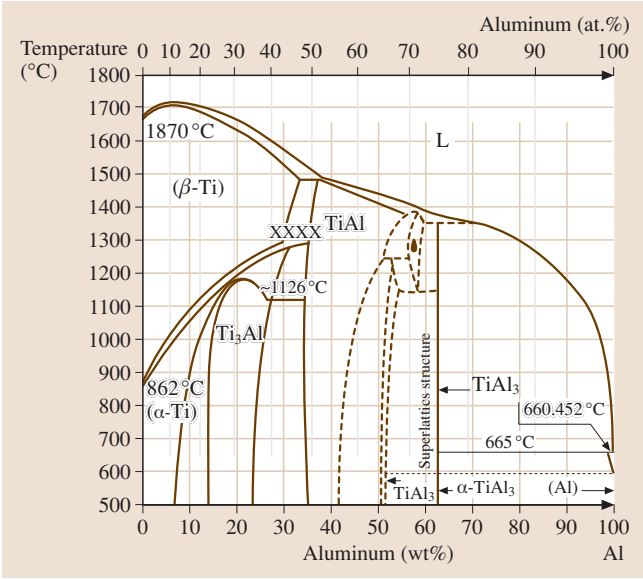


Fig. 7.307 Titanium-aluminum equilibrium phase diagram (after [7.403])

Given the short duration of the process, the temperatures necessary for the formation of intermetallic phases through atomic diffusion cannot be other than extremely high. In fact, melting provides the only explanation for the results in Figs. 7.304–7.306. Using the phase diagram of Fig. 7.307, it can be seen that the titanium-rich phase identified in Fig. 7.306 has a composition matching TiAl. Note that TiAl forms directly from the liquid phase near 1350 °C; a characteristic that strongly suggests very high interfacial heating. Similarly, the aluminum-rich phase found in Fig. 7.306 may be identified as $TiAl_3$ in the phase diagram of Fig. 7.307. This phase also possesses a particularly high melting temperature. The strongest evidence that melting occurs in magnetic pulse welds made with high collision velocities is shown in Fig. 7.308. In Fig. 7.308a are presented

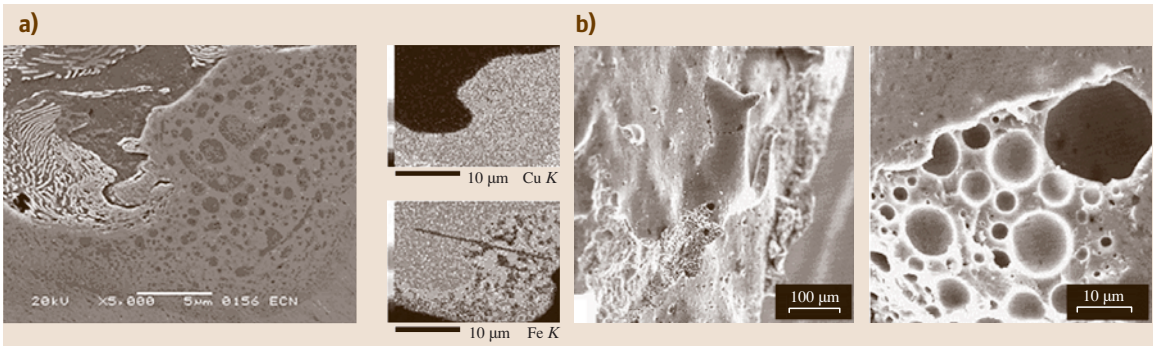


Fig. 7.308 **(a)** Secondary electron image of a low carbon steel-copper weld interface with compositional maps. **(b)** Images at the aluminum side of an aluminum-copper weld.

a secondary electron image of the weld interface between a low-carbon steel and pure copper, as well as corresponding compositional maps for aluminum and iron. Note the presence of a lamellar constituent in the upper left of Fig. 7.308a. This lamellar constituent is known as pearlite [7.405]. Pearlite is a mixture of two equilibrium phases, ferrite and cementite (e.g., Fe_3C), that both form during the slow cooling of low-carbon steel from the decomposition of the high-temperature austenite phase (another stable phase of iron, but found only at high temperatures in the absence of substantial alloying elements). In Fig. 7.308a, the pearlite is also found to disappear in the weld region, a proof that pearlite has been dissolved by a combination of high temperatures and rapid cooling. In addition to the disappearance of pearlite, dark clusters can be detected at what appear to be random locations. As validated by the compositional maps, these clusters are, as seen earlier, also evidence of the presence of new phases. Unlike previous maps, however, the results of Fig. 7.308a can hardly be explained by solid-state diffusion. Diffusion in the solid state would not produce this seemingly random clustering, but a layering of phases. The compositional maps of Fig. 7.308a reveal that the two materials, steel and copper, melted and strongly resisted mixing, as would be seen when oil and water coexist in the same medium. The two images of Fig. 7.308b present additional evidence of the occurrence of melting. In Fig. 7.308b are shown microvoids that were encountered in aluminum-copper welds. These voids were found exclusively in the regions with thick interfacial microconstituents. Once again, melting provides the sole explanation for the array of spherical voids seen in Fig. 7.308b. For the great majority of metallic materials, solidification is accompanied by a volumetric shrinkage. These voids are associated with the volumetric contraction of a new liquid phase in between the aluminum and copper workpieces.

In typical magnetic pulse welds between dissimilar materials [7.388, 393–395, 398], the intermetallic microconstituents are harder than any of the two original materials, despite adjacent strain hardening or fine-grain recrystallization that often occurs in the softest material [7.396]. Figure 7.309 shows for the aluminum-titanium weld of Fig. 7.304 that the Vickers hardness number of the microconstituent is ≈ 100. The hardness of the aluminum (initially 30) is also considerably changed. Although not represented, hardness changes in the stronger and harder titanium were unnotice-

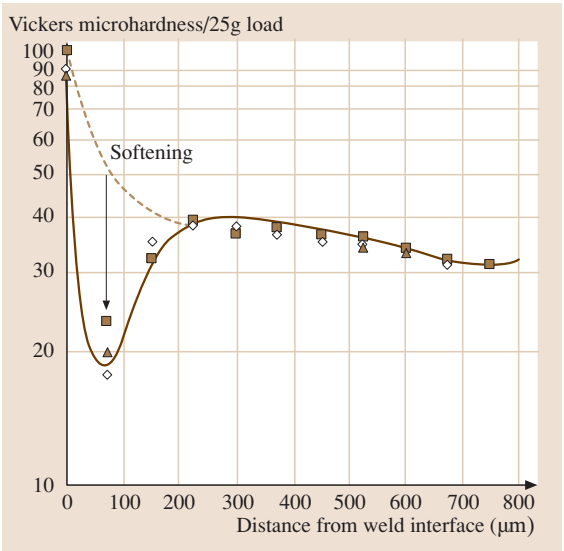


Fig. 7.309 Microhardness profile in the aluminum, as measured normal to the interface between aluminum and titanium (log scale for the microhardness)

able. These microhardness changes in the aluminum workpiece can be correlated to microstructural transformations. Because hardness variations were not monotonic, as could have been expected by strain hardening (i. e., materials strengthened as they were deformed), two competing mechanisms acted in the vicinity of the weld interface. Strain hardening clearly occurred further away from the interface, where hardness gradually increased closer to the weld. In sections of the welds where interfacial microconstituents were not seen, the hardening of the aluminum was also greater, and hardness variation was monotonic all through the aluminum wall thickness. Hardening is also greatest when the intermetallic constituents are thin [7.395]. The softening immediately adjacent to the interface, also observed in aluminum-copper welds [7.393–395], indicates that the interfacial heating was simply enough to oppose the hardening due to plastic strains. Figure 7.309 shows that the aluminum softening near the microconstituent could be as high as 30 kg/mm^2 (see arrow in Fig. 7.309). As depicted in Fig. 7.309, this decrease in hardness therefore appears to represent as much as 50% of the material initial hardness (thus strength).

While eliminating totally the intermetallic phases may reveal difficult in the case of dissimilar joints treated here, magnetic pulse welding more than any other process allows to reduce their thickness. Strengths, especially in the tensile-shear mode of aluminum-copper, aluminum-steel, and aluminum-titanium joints – among other dissimilar-metal joints – have proven to be exceptionally high [7.395]. Welds are often so strong that no other process can outperform MPW. In fact, magnetic pulse welds are often associated with weight reduction, in part because the overlaps do not have to be as long as with other processes (e.g., brazing, adhesive bonding). Since the moving workpiece is usually thin and only accounts for a small cross-sectional area relative to the area associated with the overlap, this workpiece normally fails, in many occasions leaving the joint undisturbed during testing. In applications involving tube-to-tube joints, where fluids flow on the inside of the assembly, resistance to high internal pressures is primary. A variety of tests, including hydrostatic pressure tests and helium leak tests, may be applied to measure weld quality. In some cases (e.g., heat exchanger tubes), weld joints can also be subjected to thermal cycles within a wide range of compositions at the same time they are tested for leaks. Frequently, the quality of magnetic pulse welds is proven, and magnetic pulse welding provides a hard-to-beat cost-effective engineering solution to large-scale manufacture of tubular structures.

Conclusions and Summary

MPW is a suitable process for dissimilar materials, particularly metals (including alloys) since they are electrically conductive and Eddy currents can flow in them. The moving workpiece, i. e., the one with Eddy currents, collides with the stationary workpiece with such force that both mechanical interlocking and atomic interactions occur between the two materials. Surfaces of the workpieces do not have to be cleared of surface contaminants because of the phenomenon of jetting. Magnetic pulse weld interfaces are similar to those produced by explosion welding and their properties mainly depend on the collision velocity and the collision angle. MPW possesses numerous advantages and a few disadvantages.

Several advantages of MPW are largely tied to the fact that the process can be electrically controlled from the power supply. The parameters are therefore easily controlled, adjusted, and set. This results in a highly repeatable process, which, because of the rapid discharge, is extremely fast. Other advantages of the MPW process are linked to the fact that the process is energy efficient and melting is either absent or localized to a narrow skin. As a result, the materials are not heat affected dramatically, i. e., annealed, oxidized, and residual stresses are deeply reduced compared to many other processes, as are the intermetallic phases that form between dissimilar metals.

The disadvantages of MPW are mainly related to the cost and workpiece geometry. MPW is initially more expensive than other types of welding technologies (small machines begin around $100 000), but, once up and running, it has a much lower cost than most other processes. The coils are extremely important and for optimal process performance have to be specially designed and manufactured for each application. In addition, there are high voltage and current levels in the operation of the machine that may present a safety hazard when working on the machine. Coaxial positioning of the parts to be welded is also critical, as is the angle of impact and gap. Efforts are under way to refine the MPW process and extend the process for the joining of noncylindrical components. Nonclosed coils are being developed to allow increasing part accessibility.

7.5 Rapid Prototyping and Advanced Manufacturing

This section presents basic technologies of rapid prototyping and their application in the development and functional verification of market products. Time to market is a very important factor determining the final success. The verification of virtual prototypes, while significant for the project executors, is not always sufficient for evaluation of a new product by future customers. An important argument for creating a physical prototype is the possibility of carrying out complex investigations on a tangible object. In order to accelerate the product development and evaluation, a number of so-called rapid prototyping (RP) technologies have been developed. They are also more generally called time compression technologies (TCT). A characteristic feature of RP technologies is their *additivity* – a physical object is built by *adding* material, usually in the form of layers (slices), instead of *subtracting*, as is the case in traditional manufacturing. An exception, sometimes also classified as an RP technology, is high-speed cutting (HSC) or high-speed machining (HSM).

Like virtual prototyping methods, RP technologies require a complete geometric computer model of a 3-D object to be manufactured. Various materials can be used as the construction material, e.g., photopolymers, thermoplastics, plastic films, paper and organic, ceramic or metallic powders. The material applied determines (affects, influences) mechanical and aesthetic properties of created models.

The group of TCTs also includes rapid tooling techniques that allow for building a tool to manufacture a short series of a new product (from 5 up to 100 pieces) and reverse engineering methods that allow for digitizing the geometry of an existing object, which is then processed in a CAD environment as (part of) the design of a new product.

Parts manufactured with rapid prototyping technologies still need postprocessing with specialized machining technologies for manufacturing on micro- and nanoscales. These are often called advanced manufacturing technologies.

RP technologies are especially important in developing market products since their life cycle is getting shorter and shorter and demand for new products and market competitiveness is still increasing.

Manufacture in a Competitive Market

The purpose of every manufacturer's technological activity is to process materials by means of various technologies in such a way that as a result of specific processes they become marketable products. The business purpose, however, is to configure those processes in such a way that the income resulting from the sale of the manufactured products, aimed at covering production costs and supporting design of new products and new technologies, is as high as possible. The new concepts usually have an increasing, cyclically changing nature, i. e., a step forward in technological growth happens when existing technologies and manufacturing processes do not assure superiority or at least a technological balance of a manufacturer in the marketplace.

The innovative technologies and organisational production aspects are considered to be the most important development factors of present-day production enterprises. This is especially important in long-term perspective, which usually means perfect prosperity of a company, as well as further development of innovative technologies and products. Innovative technologies create the most important indices of competitive production, namely [7.406, 407]:

- Cost reduction up to 70%
- Quality improvement up to 25%
- Increase in production flexibility up to 89.5%
- Product innovativeness up to 100%
- Technology innovativeness up to 70.6%
- Productivity increase and product line extension up to 64.7%
- Improvement of external economical indexes effect up to 44.4%
- Penetration of the international market up to 58.8%

The above data make a sufficient argument for using the newest methods, technologies, and tools in all production types, irrespective of the production scale and the company size. This is all the more important as it is possible to determine a great many variability indices of characteristic features and criteria of manufacturing systems [7.408] (Fig. 7.310).

They can be subdivided into several categories: technical, organizational, market-related, cost-related, local, and global. Of course, different criteria may also be formulated for classifying conditions, requirements, and development tendencies related to manufacturing systems. The tendencies of change will be different for different manufacturing areas, company sizes, environments, markets, or product lines.

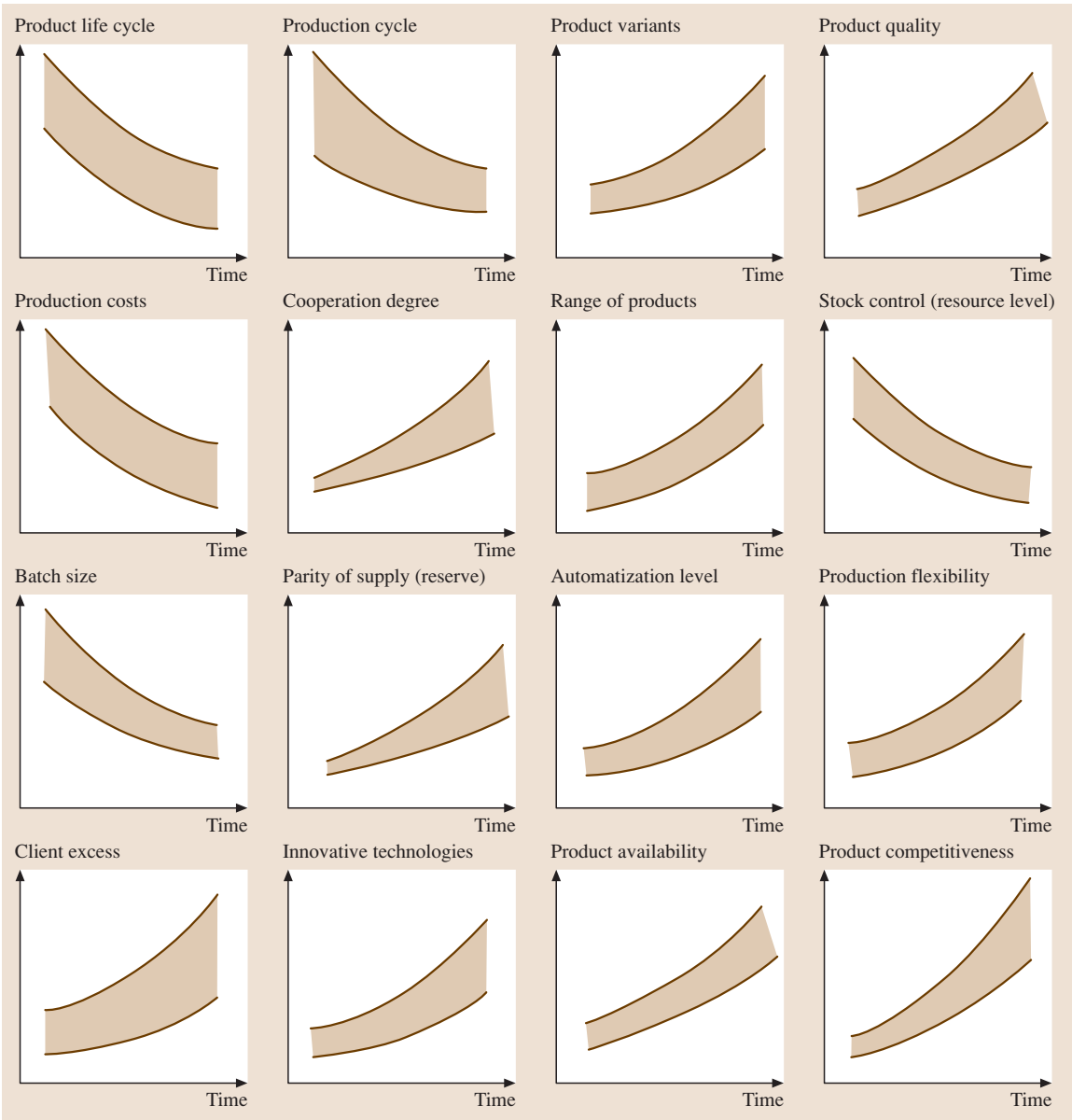


Fig. 7.310 Changes occurring in manufacturing systems

7.5.1 Product Life Cycle

The product life cycle should be defined with its determined technical, aesthetic, usability, cultural, environmental, and other functions that, as a result of operation or the passage of time, can be treated as stable, i. e., those that have not lost the value measures and criteria assigned to them as early as at the preliminary development stages.

The issue of methodological description of the product life cycle can be reduced to two important areas:

1. The basic phases of development:

- Marketing design that covers the most important features of a future product, laying the groundwork for marketing research to define its functional features with the customer's participation.

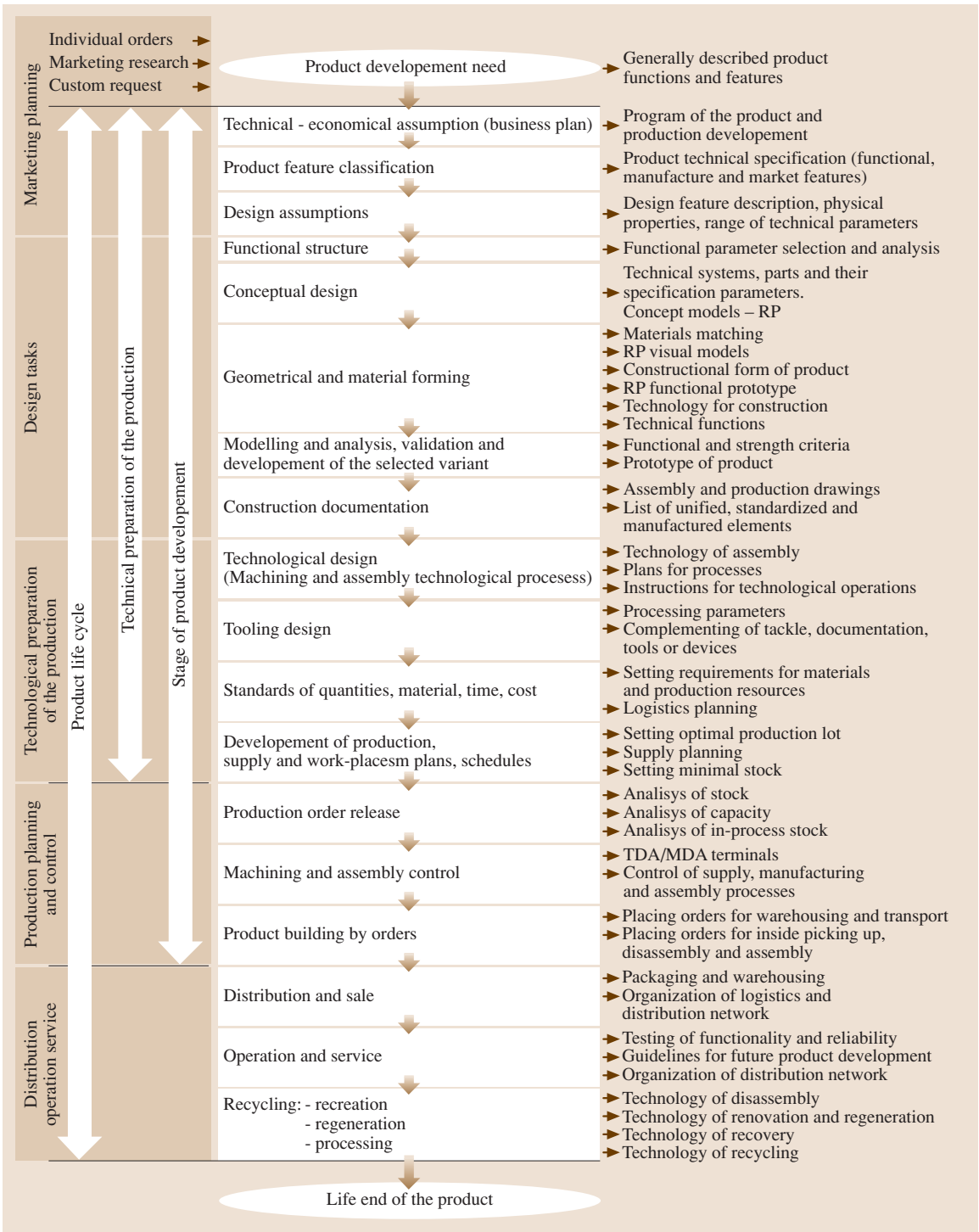


Fig. 7.311 Areas and phases of product lifecycle (after [7.408])

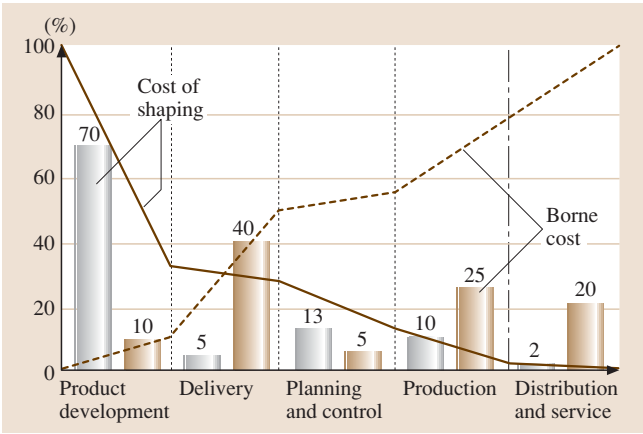


Fig. 7.312 Determined costs and costs borne at various stages of product lifecycle

- Constructional design that covers giving functions and constructional form to the new product, with special regard to geometric and material forming;
- Technological design aimed at determining technological processes together with their parameters and necessary auxiliary processes;
- Planning of manufacturing processes, considering normative planning of necessary resources and requested realization times as well as securing the production infrastructure.

2. The basic phases of manufacture and operation:

- Processing and forming of the material to the features of the product or its components;
- Assembling components and finished products manufactured to an individual customer's needs or to stock;
- Operation, servicing, and product recycling.

The basic phases of the product life cycle are shown in Fig. 7.311.

For realization and evaluation of each product's life cycle, very important are the evaluation criteria. They express not only the degree of fulfilment of functional, material-related, geometrical, ergonomic, and other requirements, but they also imply specific technological solutions. So it is a continuously challenging and inventive development and implementation process. In large enterprises, this means continuous development of one's own products, technologies, and production processes.

Costs in the Product Life Cycle

The cost of making a product is a very important factor that determines the product's price and thus its position in the marketplace. In any enterprise the development of a new product is especially dependent on the imposed functional and technical-economic requirements, formulated evaluation criteria, and market conditions. Research shows that in a product development the highest attention should be paid to the preliminary, conceptional assumptions that define the above-mentioned requirements. During new-product development, all technical aspects are determined, including the material and geometry, which in turn determine technological and process-related solutions. This proves that solutions accepted at this stage (can) influence the product's manufacturing costs up to 70% (Fig. 7.312) [7.410]. For mass-produced market products this is therefore a crucial stage since the enterprise's profitability and market position can be significantly influenced here.

Time in the Product Development

Of the general triad of antagonistic conditions and requirements imposed on today's manufacturing systems, besides cost and quality, it is time that becomes the most important. This parameter is strictly connected with flexible reactions to market demands and is a very strong factor in competitiveness, especially in distributed manufacturing systems. Globalization, using economies of scale in the manufacturing process, has contributed to the increase in production lots and distributed manufacturing, as well as forced a radical reduction in production cycles. An example is shown in Fig. 7.313.

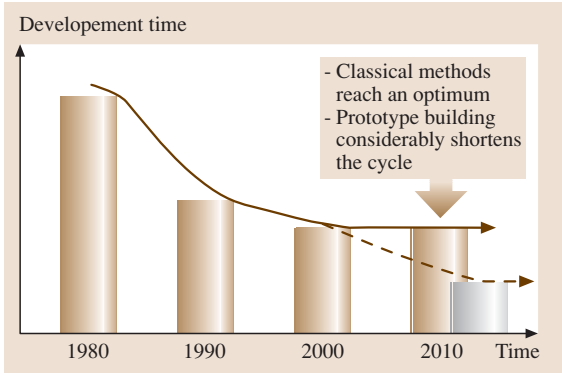


Fig. 7.313 Tendencies of reducing the product development time (after [7.409])

It is estimated that the year 2000 was a turning point for classical methods of product development and creation of new product prototypes. The market pressure on new products and more and more frequent customization require new methods and technologies. One should expect a dynamic technological development that can be seen even in rapid prototyping technologies.

The most susceptible to those changes are basic market sectors that include innovative products with a high level of application of mechatronics and IT systems as well as automotive and machine building industries. However, the contribution of information exchange and communication has increased significantly as a result of introducing the latest organizational solutions in design work – namely, concurrent engineering and collaborative engineering (simultaneous engineering). In such an organization of design work, designers spend a significant portion of their time on data search and interchange. A common concept has become that of *e-manufacturing* – completely integrated and synchronized design, manufacture, project management, etc..

Objectives of Technological Development in Manufacturing

The global economy has brought rapid development to many fields by finding new markets and using economies of scale that allow for a reduction in production costs (Fig. 7.314).

This is the current model of global objective realization by global manufacturers. The model is stable enough but not very innovative. The purpose of all manufacturers is to introduce their products to zones 3 and 4, which means dynamic technological development resulting in introducing to current and new markets new products representing new functions and usable features and generating the highest added value with not necessarily large production volume.

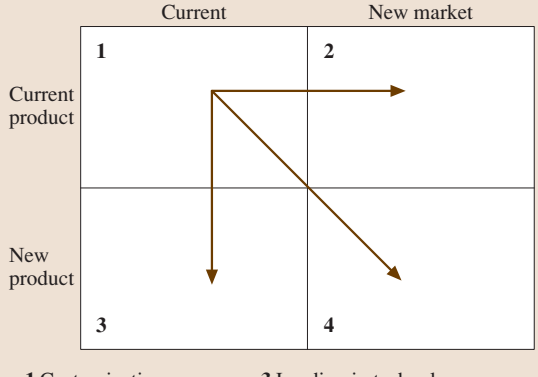


Fig. 7.314 Objectives of development in manufacture (after [7.411])

7.5.2 Rapid Prototyping Technologies

An important stage of new product development is the manufacture and evaluation of its prototype. In many applications a virtual model does not allow for a full verification of a new product and it is necessary to make a physical prototype, e.g., using a technology from a group called rapid prototyping (RP), also called more generally time compression technologies (TCT).

Rapid technologies usually form a smooth transition in a product's development (cycle) between a virtual model and the series production [7.412–414]. Mechanical properties of RP models are, from the point of view of their applications in subsequent stages of product development, important features that are the prerequisite of effectiveness of a given method for designing properties of the planned product.

Basic Knowledge on RP Technologies

A characteristic feature of RP technologies is a shorter time to obtain fully valuable prototypes of new products or their components owing to the elimination of tools required in traditional mechanical technologies (dies in casting, cutters in machining, electrodes in spark erosion, etc.). Another feature is that the processes are based on 3-D computer models of the parts to be manufactured [7.408, 412–416].

The third attribute of RP technologies is their *additivity* – a physical object is built by *adding* material, usually layerwise, instead of *subtracting* it, as in traditional manufacturing. Depending on the RP technology, prototypes are manufactured by an additive method in computer-controlled machines and can be built from paper laminated object manufacturing (LOM), laser-hardened resin (stereolithography), plaster powder (3-D printing), sintered powders of plastics or metals (selective laser sintering), etc. (Fig. 7.315).

As an exception to building with a laminar growth of material there are technologies of HSC or HSM, sometimes also classified in the RP group, where a laminar loss of constructional material is realized by machining. The use of modern tool materials and designs, the progress in machining technology, and advanced constructional solutions in high-speed milling machines allow for obtaining high productivity and very high

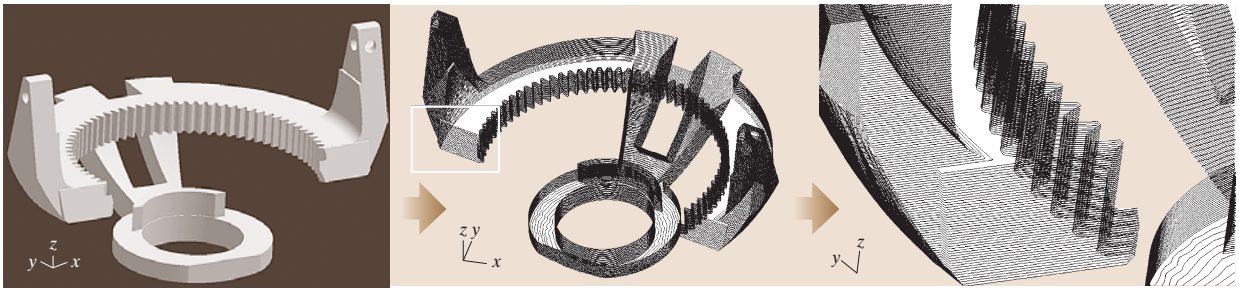


Fig. 7.315 Idea of slicing process in layer additive technologies

machining precision in processing metals and other materials. Therefore, the manufactured objects can be pattern models, prototypes, tools for shaping (dies, stamps), molds for injection forming of plastics, or molds for metal casting. Direct application of HSM for manufacture of finished products in one-off or series production is an example of so-called rapid manufacturing (RM).

Most frequently, the selection of a traditional, an RP, or HSM technology is determined by the time and costs of manufacturing a physical model (a pattern or a prototype). These quantities are shown in Fig. 7.316 [7.417].

The technology of additive manufacture of physical objects is more effective than machining (realized by the HSM or the more conventional milling on NC machine tools). The time and cost of additive manufacturing are only slightly dependent on a model's degree of geometrical complexity. The values shown above will soon change as a result of quick development of both methods. During the last few years, for example, the rate of physical model creation with RP methods has increased several dozen times due to the application of:

- Multiplying the number of laser systems building the same object in a machine
- Higher power lasers allowing faster scanning of a layer being built
- Faster drives and control systems of laser beam deflection systems
- New organization of the manufacturing process

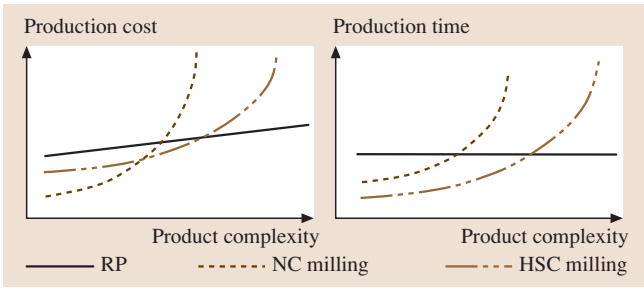


Fig. 7.316 Comparison of time and production costs of prototype models made by the RP and RM-HSC methods

Additive Technologies in Rapid Prototyping

RP describes the generating methods used for laminar creation of components and prototypes with a high degree of geometric complexity. Various technologies are based on one basic principle: 3-D geometrical models created in 3-D CAD systems are *sliced* into layers and in this way reduced to a 2-D form. Such prepared *flat* objects are bonded together to create a 3-D physical model. In many RP methods materials are spot hardened with laser beams (photochemical methods). This process is repeated for all layers of the object being created. Alternative methods are based on laminar cutting out of the contours with a laser beam or on bonding the powdered material with a binder into laminar structures. Since implementing stereolithography as the first commercial RP method in 1987, many other methods have been developed that use the same principle of laminar object structure [7.412, 418, 419]. Classification of RP methods based on applied and processed materials is shown in Fig. 7.317 [7.420].

The necessary condition and the starting point for application of all RP methods is the existence of a complete 3-D geometrical description of the component being prepared. The object geometry, described in a CAD system, is simplified for further mathematical processing by approximating the object's surfaces with triangle meshes and transformed into a data exchange format standard for most RP methods – stereolithography language (STL). Very important is setting the proper number of triangulation parameters (*chord height, angle control*), since they largely determine the quality of the obtained approximation. The maximum

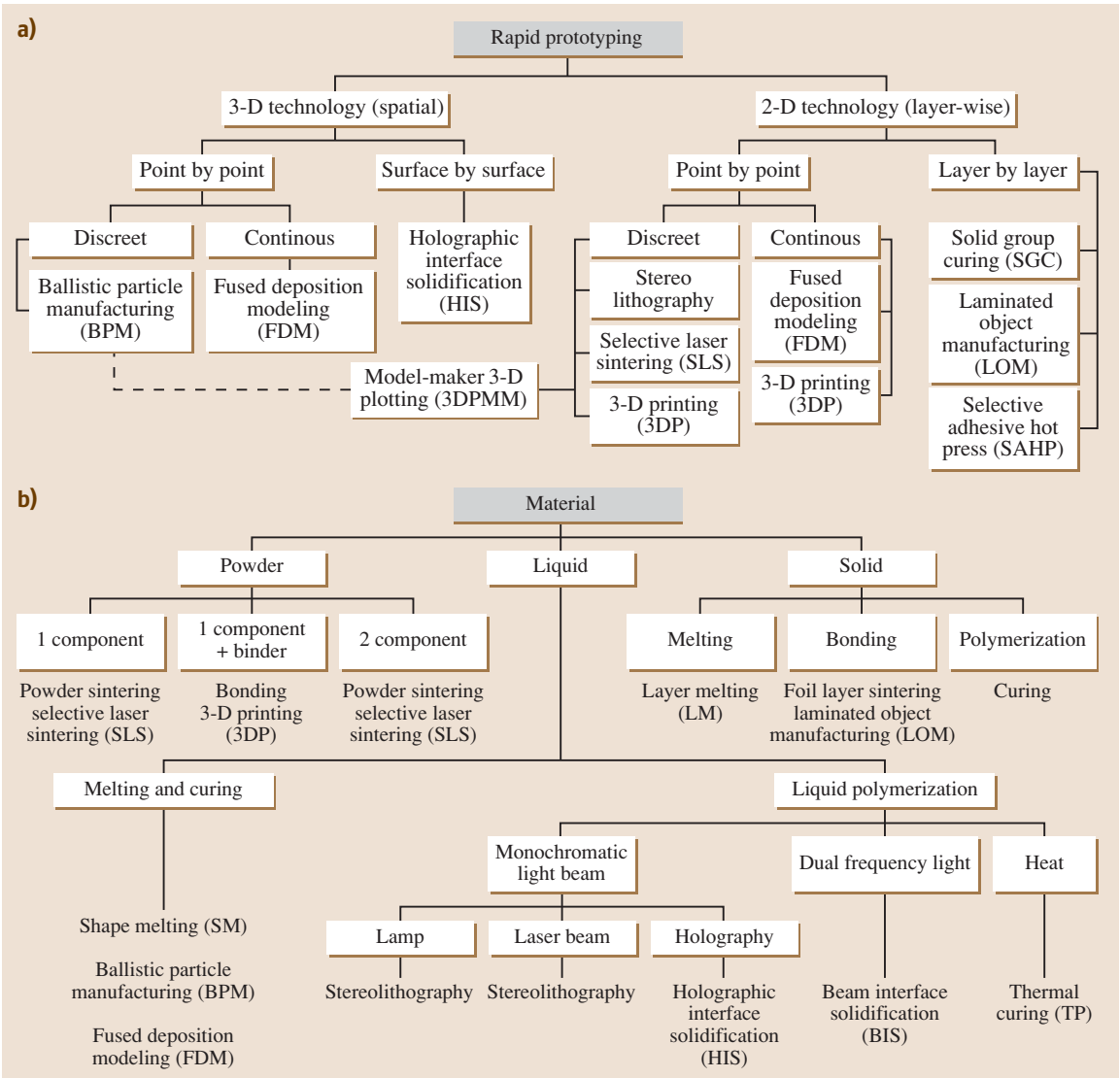


Fig. 7.317a,b Classification of RP methods based on building method **(a)** and based on applied materials and processes **(b)**

permissible chord error (in millimeters) is set by the *chord height* value, and the *angle control* value determines the maximum permissible angle between two triangles. The STL data of an object are further processed in such a way that the 3-D geometry is broken into individual sections (layers) with a defined height (SLI format). Usually, those layers are 0.1–0.2 mm thick.

Classification of RP Technologies. The wide range of solutions in the field of incremental laminar technologies, also referred to as *layer manufacturing* (LM), uses three basic principles of bonding material: chemical bonding, sintering, and gluing. They also include materials in the form of epoxy and acrylic resins, metal and plastic powders, ceramics, and paper. The main laminar methods are shown in Fig. 7.318 [7.421].

Abbreviations (names of developers in brackets): FDM – fused deposition modeling (Stratasys), CC – contour crafting (University of Southern California), SMM – Sanders model maker (Sanders prototype), DMD – direct metal deposition (University of Michi-

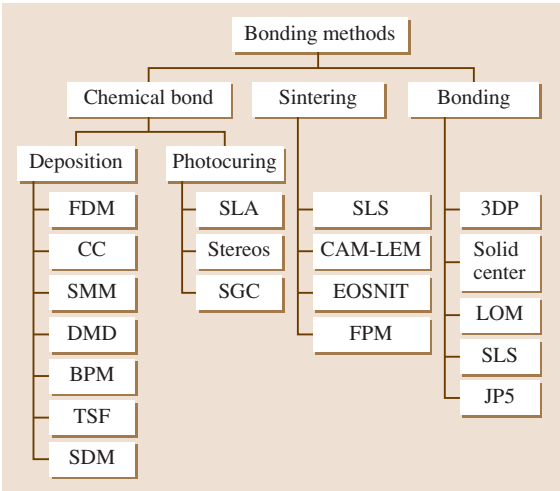


Fig. 7.318 Classification of RP technologies based on the commercial names

gan), BPM – ballistic particle manufacturing (BPM, Inc.), TSF – topographic shell fabrication (Formus), SDM – shape deposition manufacturing (Stanford University), SLA – stereolithography (3-D Systems, Inc.), SGC – solid ground curing (Cubital), SLS – selective laser sintering (University of Texas and DTM), FPM – freeform powder molding (RPI), 3DP – 3-D printing (Massachusetts Institute of Technology and Z-Corp.), LOM – laminated object manufacturing (Helisys), JP5 – JP5 system (Schroff), CAM-LEM – computer aided manufacturing of laminated engineering materials (Case Western Reserve University and CAM-LEM Inc.), EOSINT – laser sintering (EOS GmbH), STEREOS – laser photolithography (EOS GmbH), solid center – paper lamination (Kira Corp.)

Some of the methods, like stereolithography, require adding to the model some supporting structures to enable removing the prepared part from the supporting platform in the machine, and to protect the model against deformation during the building process. The basic advantages and disadvantages of RP techniques are given in Table 7.56 [7.412].

Models and prototypes made with RP techniques can be classified in the following way (slightly modified definitions given in [7.412]):

- The *conceptual model* describes the main geometric and dimensional proportions in a simplified way that allows clear and persuasive presentation of the design solution idea. Detail minuteness of the solution – low.
- The *ergonomic model* determines the boundary conditions of the solution with respect to safety and comfort of the product operation by the future user. Detail minuteness of the solution – medium.
- The *geometrical model* fully reflects the geometrical features of the 3-D CAD model. Often called a visual prototype (model). Detail minuteness of the solution – high.
- The *constructional model* is a synthesis of the above-mentioned models. Detail minuteness of the solution – high.
- The *functional model* has features of the constructional model but also allows for testing performance and basic functions of a product. It is the final stage

Table 7.56 Main advantages and disadvantages of the RP techniques

Advantages	Disadvantages
• Rapid creation of physical patterns • Pattern component at the disposal as early as during the structure development • Especially suitable for – parts with complex geometry (first of all outlines), – free-shaped surfaces • Low manufacturing costs in comparison to other methods (milling, turning, electromachining etc.), mainly for small number of pieces • Possible application of different methods within the whole chain of processes (rapid engineering)	• Limited dimensions of the created objects • Limited range of materials • Components meet the requirements to a limited extent only • Limited exactness (ca. ±0.1 mm), surface quality conditioned by the applied manufacturing technique • Additional smoothing often necessary

of the product design. Detail minuteness of the solution – high.

- A *functional prototype* permits evaluation of the main functions of the solution in close-to-reality conditions, with limited operational parameters.
- A *technical prototype* has all the functionality and aesthetic features of a mass product that allow subjecting it to examination and evaluation within the whole range of operational parameters. It is used for examination and determination of allowed operational parameters. After evaluation by potential users and possible corrections (usually in ergonomic and functional models), it is moved on to series production.

The classification given above is not explicit and, depending on the type of product, features of some models can be synthesized or be absent entirely. They will be understood in different ways in manufacturing processes of cars, home appliances, TV sets, table lamps, or perfume bottles.

Typical application areas of these techniques are:

- Design and ergonomic studies
- Examination and evaluation of design solutions on the basis of physical models and research methods from the scope of photoelasticity, thermovision, X-ray radiography, flow modeling, etc.
- Analysis and evaluation of manufacturing processes, especially assembly processes
- Examination and modeling of flows in plastics forming
- Marketing examination and evaluation of new products
- Testing multifunctional models in casting and plastic working
- Modeling and manufacture of osseous and *soft* implants in medicine

Owing to the application of these methods it is possible to significantly reduce the product life cycle as well as reduce the costs and risk of its development and implementation. The possibility of manufacturing objects without any special tools, molds, or dies has undoubtedly become the decisive factor in the increasing interest in these methods to minimize investment risk. The range of materials used in RP technologies is still growing and includes metals, polymers, ceramics, timber, fiber-reinforced materials, and various metal- or polymer-matrix composites.

In RP processes some problems occur related to the quality of the obtained objects. Besides the stepwise appearance of inclined surfaces, resulting from laminar object preparation, there are also problems related to material shrinkage during processing (e.g., in stereolithography) and with porosity (SLS). Therefore, efforts are being made to develop materials with lower shrinkage and to formulate suitable strategies of manufacturing processes [7.413, 415, 416, 422–424].

RP Technology Application Areas. RP technologies are especially useful in these industrial sectors and these fields where it is necessary to create physical models and respond quickly to market demands. The main application areas of the RP techniques are as follows [7.412]:

- Prototype building for:
  - Verification of design solutions
  - Analysis and evaluation of design solutions
  - Examination of flows
  - Research in wind tunnels
  - Selection of construction materials
- Physical model building for:
  - Searching for design solution ideas
  - Building design and industrial design
  - Marketing presentations for customers
  - Problem solving by the *case study* technique
- Manufacture of components for:
  - Production of tooling and accessories
  - Production of auxiliary means of production
  - Marketing research with a trial lot
- Design and manufacture of tooling for:
  - Planning of production processes, especially assembly processes
  - Design and manufacture of prototype tooling, especially for sheet metal forming
- Design and manufacture of patterns and models for:
  - Casting technologies, including sand casting and lost-model processes
  - Vacuum forming
  - Hydro- and thermoforming
  - Forming by metal spraying on a pattern
  - Epoxy techniques and materials

As determined by research in companies using RP technologies, the most important and largest area of RP application is in the manufacture of functional prototypes subject to constructional analysis in working conditions of finished products and their manufacturability analysis. Figure 7.319 shows the shares of application areas of models made by rapid prototyping techniques, obtained on the basis of data acquired in

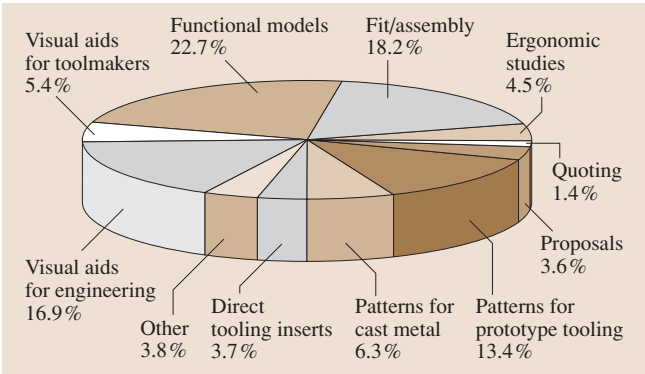


Fig. 7.319 Application of RP technology in industry

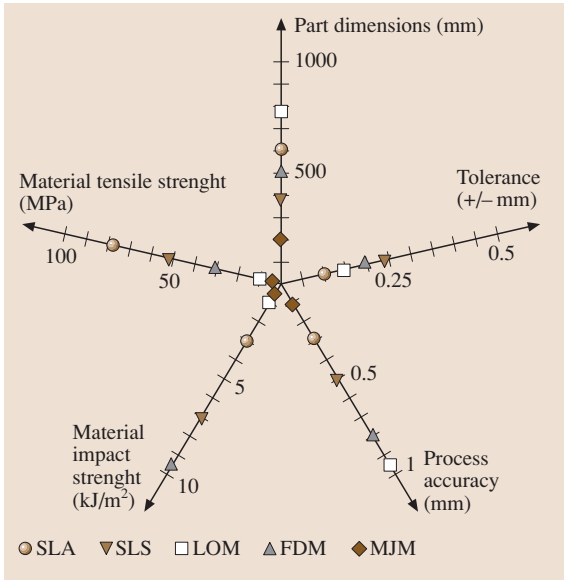


Fig. 7.320 Mechanical properties of materials applied in RP technologies (after [7.426])

2001 from 57 companies (14 users of RP systems and 43 suppliers of RP services), based on industrial data from their customers and their applications [7.423,425].

The data show that 41% of all models were used for analysis of assembly processes and functional examination. About 27% of the models were used as visual aids at the stages of design, making tools, and analyzing orders and offers. Over 23% of the RP models were used as patterns in the manufacture of prototype tooling such as casting patterns as well as for the manufacture of cores and inserts.

The application range and frequency of RP methods depends on the properties of the processed materials and their condition after process termination. Figure 7.320 illustrates the basic mechanical properties of materials processed in RP processes, durably formed in the manufactured products.

It can be seen from the above specification that the largest application field of RP models is the analysis of constructional solutions, assembly, and functional examination. These applications require that prototype models be very precise and represent geometrical features, which in turn implies usable properties of prototype components.

Review of Selected RP Technologies

Stereolithography. The oldest, most popular, and best known RP technology is *stereolithography* (SL). It was first presented by 3-D Systems Inc. in 1987 at the AUTOFACT Fair in Detroit. In this method master objects (physical models) may be manufactured without casting molds or tools. The first realization stage is geometric modeling of an object in a 3-D CAD system. At the next stage geometrical data of the object are processed by a special program that splits the 3-D model in the *X*/*Y* plane into layers usually 0.1 mm thick. As a result, a set of layers is obtained that allows for a suitable program to control a UV laser beam (He-Cd laser, with the power ranging from 20 mW to 1 W). The software controls the beam movements in such a way that by scanning it exposures a section of the 3-D model on the surface of liquid polymer. In the exposed place photopolymerization occurs, i. e., the polymer is transformed from liquid to solid state [7.427]. In this way a fragment of the physical 3-D model is created (Fig. 7.321).

The material out of which a model is made is a liquid plastic hardened (polymerized) with a laser light of a suitable wavelength. The object created in this process, immersed in liquid polymer, is gradually moved down (along the *z*-axis) in such a way that the subsequently hardened layer is connected with the previous one, making a uniform body. Before each stage of exposing a new layer, the polymer level is evened, so that the exposed layer height is identical on the whole surface. After the object creation process is terminated, the object is removed from the liquid polymer and additionally hardened by exposure to UV light. Since the part is built in a liquid environment, in some cases like long, flat, irregular objects etc., it is necessary to complete the model structure with proper supporting elements in order to make it more rigid. The time of the object surface scanning with a laser beam depends on the complexity of individual layers, the velocity of the laser beam movement, and the time of polymer harden-

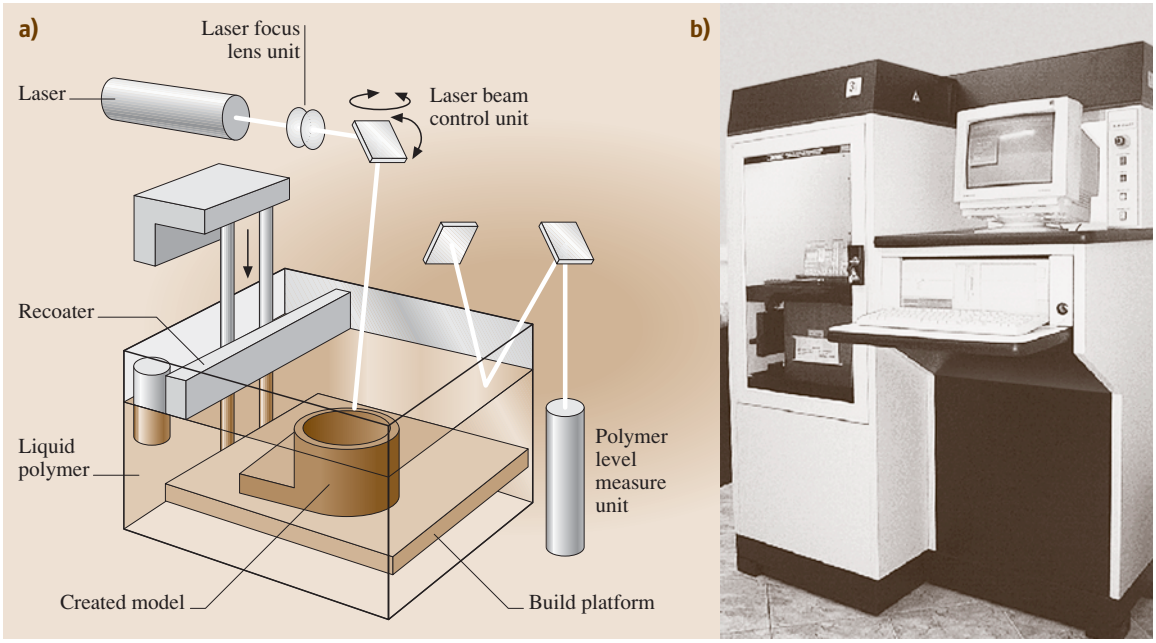


Fig. 7.321a,b Diagram of model creation by stereolithography **(a)** and a stereolithography machine SLA-250 **(b)**

ing and bonding with the preceding layer. Thus, SL is based on achievements in such fields as polymer chemistry, CAD/CAM systems, and laser technology.

The main stages of a stereolithographic model creation are:

- Building a model in a 3-D CAD system
- Exporting the model in the STL format
- Designing the position and geometry of supporting elements (new 3-D model)
- Setting the model-building parameters in a SL machine
- Dividing the 3-D model into layers according to the set parameters of the physical model creation
- Building the physical model in the photopolymerization process

At the time of building a model, a so-called *recoating* system ensures a constant thickness of the layers. The system skims needless polymer allowance from the last,

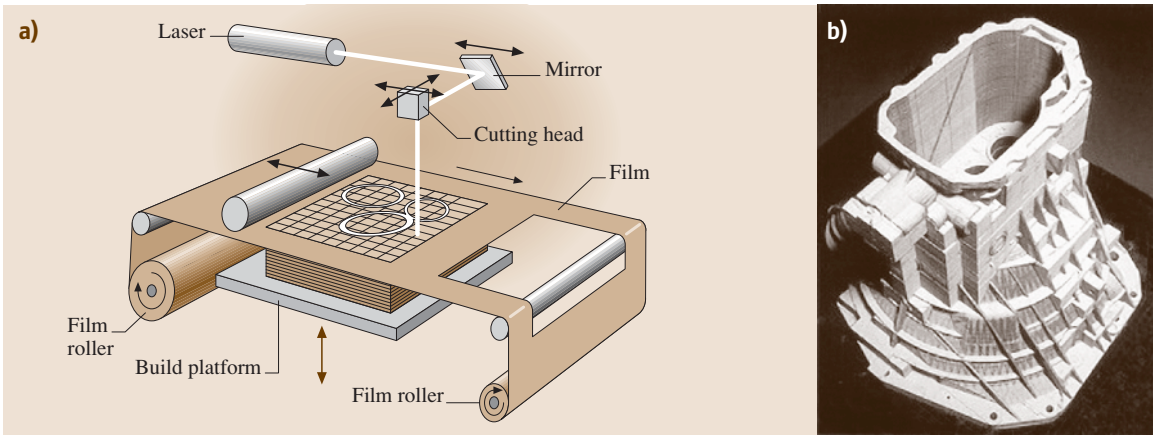


Fig. 7.322a,b Diagram of model manufacturing by LOM method **(a)** and a model prepared this way **(b)**

hardened layer of the photopolymer. When relatively large layers are scanned and hardened, it is necessary to apply an additional liquid layer to ensure constant thickness of the subsequent hardened layer.

SL is widely used in building models and prototypes of products and usable objects in many fields such as industrial design, automobile manufacturing, and home appliances, as well as medicine and architecture. To obtain specific properties of models and prototypes, some additional processes are often required that give the object suitable, required features.

Solid Ground Curing Technology. The method of direct substrate hardening *solid ground curing* (SGC) was developed by an Israeli company called Cubital Ltd. and is realized in the form of the SOLIDER system. The principle is similar to that of the SL method, but there are several significant differences [7.428].

In this case the model is also built layer by layer by hardening a photopolymer. However, the UV light source is not a laser but a UV lamp. Moreover, individual model layers are created generally by exposure of a previously prepared mask of the given layer on a glass plate. This mask is made using a technique similar to that used in a laser printer, although a negative image of a model's layer is created. This means that in the places where the object outline is to be created (exposed), the mask surface is transparent and can transmit UV light, but in other places a nontransparent toner is deposited on the plate. The glass plate, after cleaning, can be repeatedly used for mask making. The pot with the created object moves not only vertically (consecutive model layers) but also horizontally as it is necessary to perform subsequent stages of object creation on individual stations of the SGC machine. When consecutive model layers are created, the nonhardened polymer is collected and free space in the object is filled with wax. This makes it possible for the created model to stiffen and no special supporting elements are required. A cold metal plate is used for wax hardening. Each created model layer is leveled to proper height by milling, which makes it possible to *undo* operations, i. e., to cancel results of previous actions. Next, a subsequent layer of polymer is applied on a smooth and even surface of the created object.

Laminated Object Manufacturing Technology. The method of *laminated object manufacturing* (LOM) was developed by the American company HELISYS.

In this method an object is created by cutting out outlines of individual layers of a model with a laser (with power from a few dozen to a few hundreds watts) and sticking consecutive layers of a film moving by means of rollers over the model being built [7.413]. The model is located on a platform that, along with the model creation, is gradually lowered down by a thickness of consecutive model layers. The film is coated underneath with special glue and the cut-out layer is stuck to the previous one by means of a hot roll that melts the glue and presses and levels the surface of the object being created. As the thickness of the film is not exactly constant, a special sensor is used for measurement of the model height. Plastic, ceramic, and metal films can be used. To facilitate removing the excess material from the finished model, especially if it is not prismatic and has complex internal spaces, the laser beam cuts characteristic squares on the film areas not used for the model creation. After pressure welding, the squares make prisms that are easily removed from the model body. This part of the material is waste. However, this material cannot be removed from completely or partially closed internal spaces of the model. This is a disadvantage of the method that can be omitted by subdividing the model into several parts.

3-D Printing Technology. A simple and cheap method of manufacturing conceptual models is *3-D printing* (3DP), developed at the Massachusetts Institute of Technology [7.424, 429, 430]. The principle is based on laminar bonding of powdered material with a binder applied by a printing head. A diagram of the 3-D printer operation is shown in Fig. 7.323 and exemplary model in Fig. 7.324.

Models manufactured with this type printer are made of powdered starch or powdered plaster. The building process is as follows:

1. The printer applies a powder layer from a container to cover the surface on the molding platform.
2. The binder is overprinted on the prepared substrate to form the first layer of the object cross-section. In overprinted places, the powder is bonded (glued) with the binder. The remaining unbounded powder, in unchanged form, serves to support the physical model.
3. After a layer is completed, the platform with the model is slowly lowered down by a distance equal to the layer thickness.

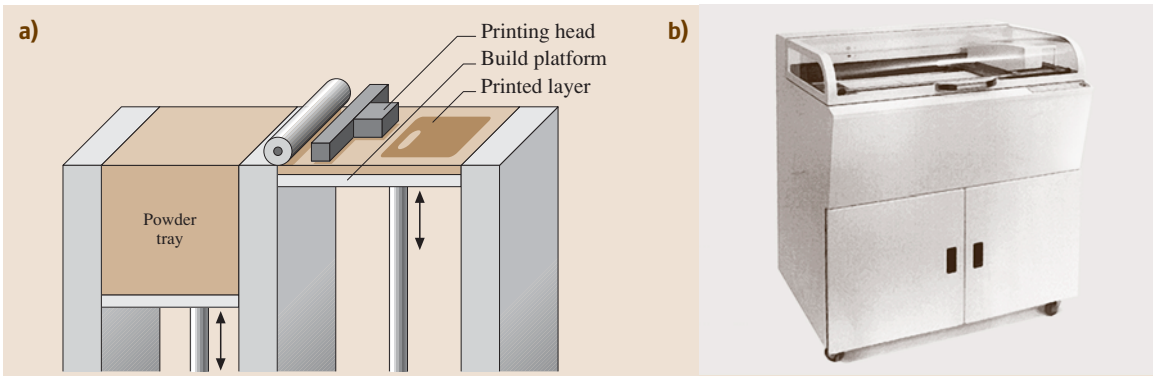


Fig. 7.323a,b Functional diagram of a 3-D printer Z400 made by Z Corporation (**a**) and view of the equipment Z400 (**b**)

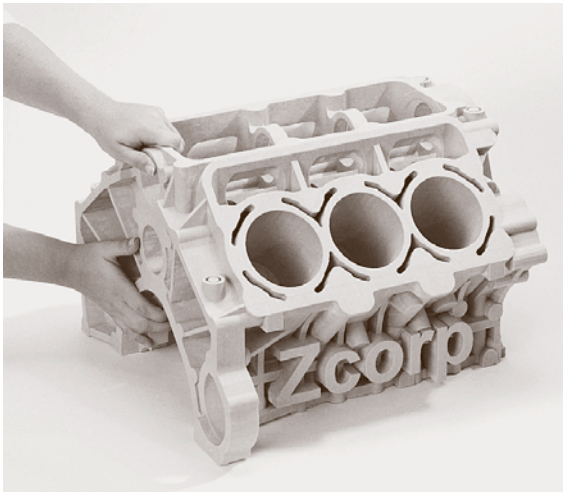


Fig. 7.324 Exemplary model made in 3DP technique

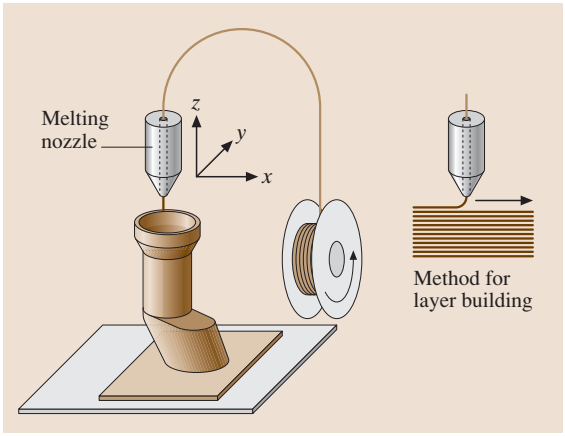


Fig. 7.325 Diagram of model manufacture by FDM technique

4. A subsequent powder layer is applied and the binder overprinting process is repeated.

This cycle is repeated until the whole physical prototype is completed. When building is terminated, the platform with the model is raised and the free, unbound powder is sucked off. The part prepared in the 3-D printer can be impregnated with various materials to increase its strength or flexibility.

Fused Deposition Modeling Technology. In the *fused deposition modeling* (FDM) method, developed by Stratasys, the physical object is built incrementally by applying layers of a material melted by means of a properly controlled nozzle [7.431]. The operating principle is shown in Fig. 7.325.

Direct Laser Sintering and Melting Technologies. A laser beam causes the sintering of determined areas of the powder layer, which sticks to the previously applied layer to provide a uniform structure of the model being created. This method, unlike the previously described RP methods where the material was transformed from the liquid into the solid state, is characterized by material transition from the solid (powder), through the liquid, and again to the solid state (sintered). The technology allows for manufacturing all kinds of models, but also high value tools for series production, like dies, stamps, casting molds, or injection molds for plastics, as well as finished products manufactured in single pieces or small lots.

Selective Laser Sintering Technology. The *selective laser sintering* (SLS) method was developed at the

University of Texas in Austin, USA. In this method a special roll spreads on the working platform a powder layer that is then locally sintered with a laser beam (with power up to several hundred watts and even several kilowatts) (Fig. 7.326).

The SLS method uses such materials as plastics, wax, metal powders (Fe-Cu), and mixtures of metallic and ceramic powders. An advantage of this method is that, like the 3DP and SGC technologies, it does not require any supporting elements when the model is asymmetrical and not too large. However, in very complex and large models it can be necessary to use such stiffening supports. At the final stage of model creation, the nonsintered powder is removed [7.427, 432].

In the SLS process metal powder grains coated by polymer binder are joined by laser power. The disadvantage of SLS is the necessity of removing the binder and infiltration of the product with bronze.

DLF/DLM Technolgies. *Direct laser fabrication* (DLF) is a new technology of rapid prototyping or even rapid manufacturing, developed at Birmingham University [7.433]. It is a modification of the SLS method; like SLS, it is based on laser sintering of consecutive powder layers, but, unlike SLS, it concerns exclusively metal powders and does not require adding polymers or resins that maintain the solidity of the created 3-D ob-

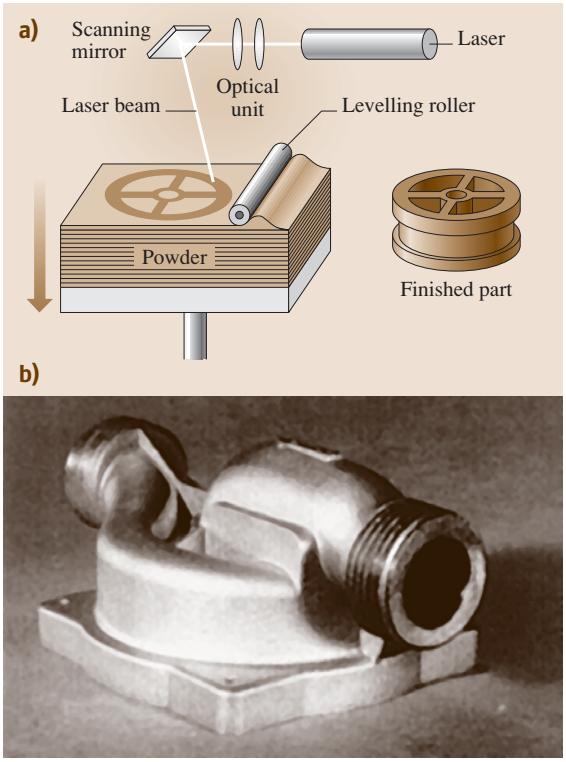


Fig. 7.326a,b Diagram of model manufacture by SLS method (a) and a model prepared this way (b)

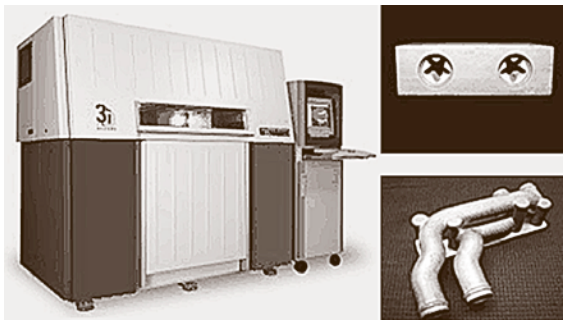


Fig. 7.327 Sinterstation HiQ System of 3-D Systems. Technical characteristics: materials: steels ST100, ST200 and A6; no full melting, infiltration required; layer thickness 70 μm; accuracy 0.05–0.1 mm; build volume 330 × 380 × 457 mm; speed 50–80 cm^3/h

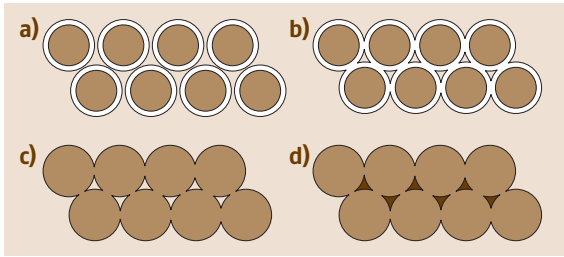


Fig. 7.328a–d SLS process: (a) Loose powder grains before sintering, (b) sintered grains (c), grains without melted binder (d), infiltrated grains

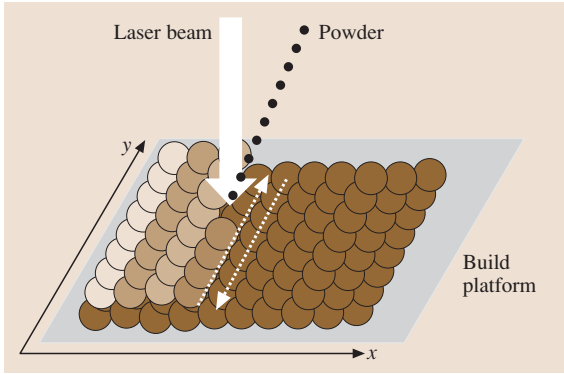


Fig. 7.329 Operating principle of DLF method

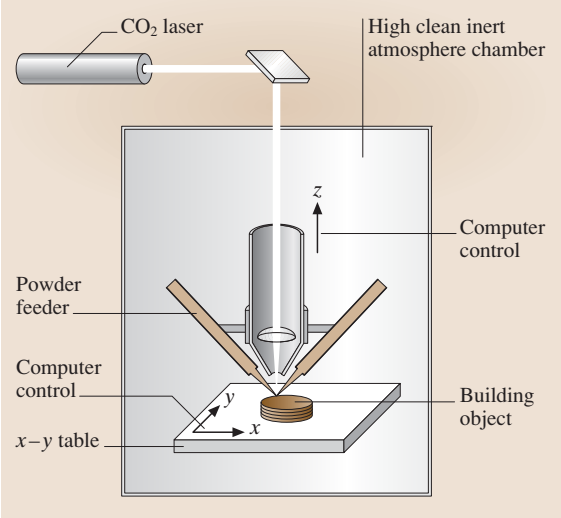


Fig. 7.330 Layout of a DLF machine

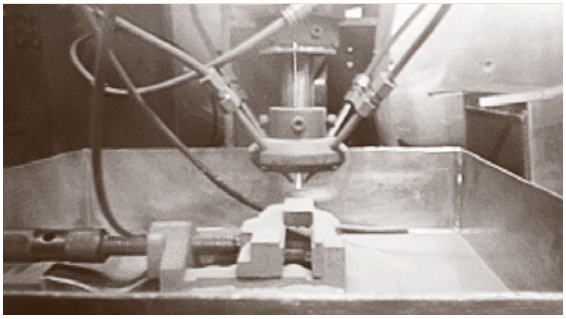


Fig. 7.331 Inside of a DLF chamber

ject and must be later baked and impregnated to reduce the material porosity.

In the DLF method the laser beam (CO_2, 200–1000 W) is focused on a determined point of the created object, which permits temperature increase in the working point up to ca. 1000 °C. A powder is also supplied to the focal point. Owing to that, the obtained object is characterized by almost full density with no necessity to infiltrate it after removal from a DLF machine.

It is possible to control the DLF process parameters in order to obtain objects with various porosities. In this way, one can manufacture objects with controlled porosity, e.g., for medical applications or for the production of filters.

The material structure of a created model is significantly affected by thermal conditions, depending on, among other things, laser power, scanning speed, and

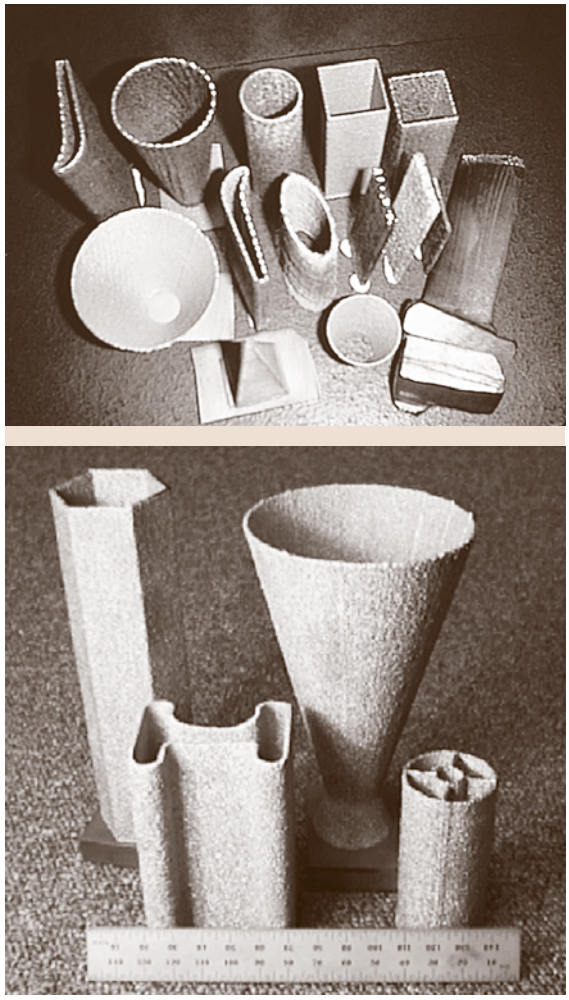


Fig. 7.332 Parts made with the DLF technology from steel and titanium alloy

the quantity and composition of the powder applied. Thanks to the *DLF* technology it is possible to manufacture easily and quickly parts of materials difficult for casting and other more traditional processing. The metal powder fused into a finished product is completely free of pores and its strength is equal to that of a solid material.

Direct Metal Laser Sintering Technology. The direct metal laser sintering (DMLS) process was developed by the EOS company for the equipment EOSINT M (Fig. 7.335) [7.434, 435]. This technology uses a mixture of bronze and nickel powders without additional bonding components. The powders are characterized by

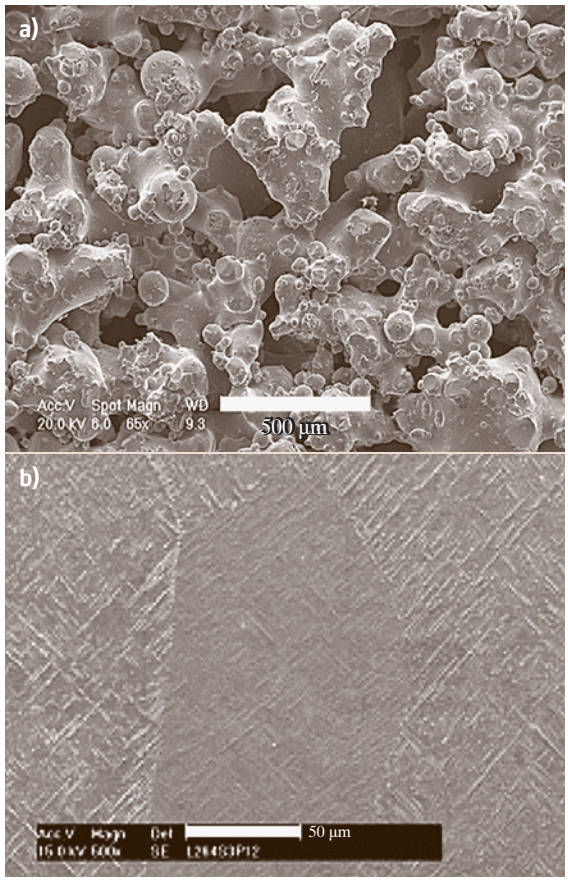


Fig. 7.333a,b Effect of laser power on model structure: laser power 180 W: **(a)** Porous material, laser power 264 W, **(b)** completely fused material

different melting temperatures, which meets the SLS principle based on the fact that the absorbed laser energy melts the low-melting phase (bronze) and wets the high-melting phase (nickel) that remains solid and develops by irreversible change of the crystal lattice. This ensures constant volume during sintering and thus results in high exactness of geometrical representation of the part. The occurring porosity within 25–45% is removed by infiltration of the mold surface with epoxy resin. The application range of this method can be expanded by use of powdered steel materials. This group includes DirectSteel 50-V1, the material developed by EOS for the equipment EOSINT M 250 Xtended. Steel powder with a grain size of 50 µm permits obtaining steel forming inserts of injection molds and thus producing series even up to 100 000 pieces. This technique

Fig. 7.334 EOSINT M270 – EOS device. Technical characteristics: materials: tool steel H20, steel DS20,Ti6AL4V, stainless steel EOS 17-4 PH, bronze DM20; full melting; layer thickness 20–100 µm; accuracy 0.05–0.1 mm; build volume 250×250×215 mm; speed 5–70 cm^3/h

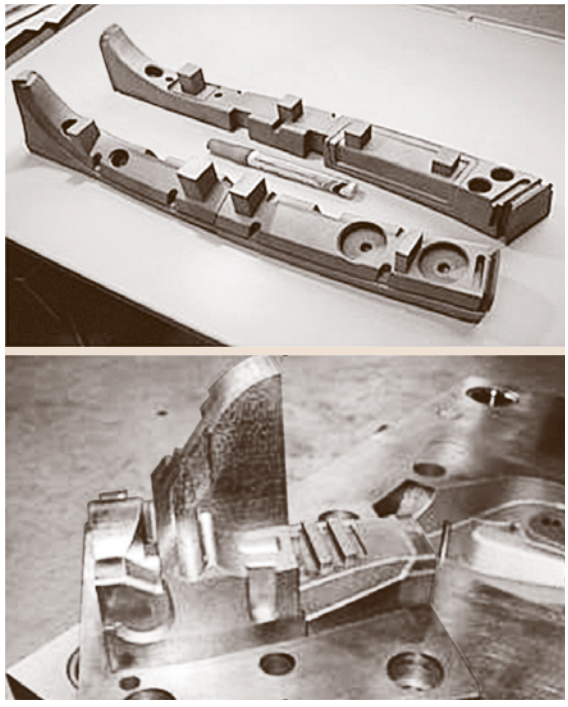


Fig. 7.335 Tools manufactured by DMLS

can also be applied for pressure die casting of highly abrasive plastics to obtain series of molding up to several hundred pieces.

SLM Technology. Since 1995 SLM technology has developed in close cooperation between F&S and the Fraunhofer Institut für Lasertechnik (ILT, Aachen). SLM is a rapid tooling method that uses commercially available one-component pure metal powders (with no binders) with a corn size of 10 to 30 µm for the production of 100% dense parts. To date powders from stainless steel, tool steel, and aluminum can be processed. Currently it is marketed by MCP [7.436, 437].

Especially by the usage of tool steel the produced SLM parts reach a new dimension with respect to maximum load and wear resistance. Right after the production process tool steel SLM parts show a hardness of ≈ 550 HV. Already in 1997 mold inserts for plastic injection molding and aluminum die casting were built and tested by industrial partners. The potential application spectrum of such parts reaches beyond usage as a tool prototype for small series up to ready-for-use series production.

Postprocessing known from the methods of SLS to increase density and strength is unnecessary in SLM.

The characteristics of the MCP Realizer II are as follows:

- Many different materials may be used to build parts from almost all metal powders (e.g., titanium, steel, Co-Cr).

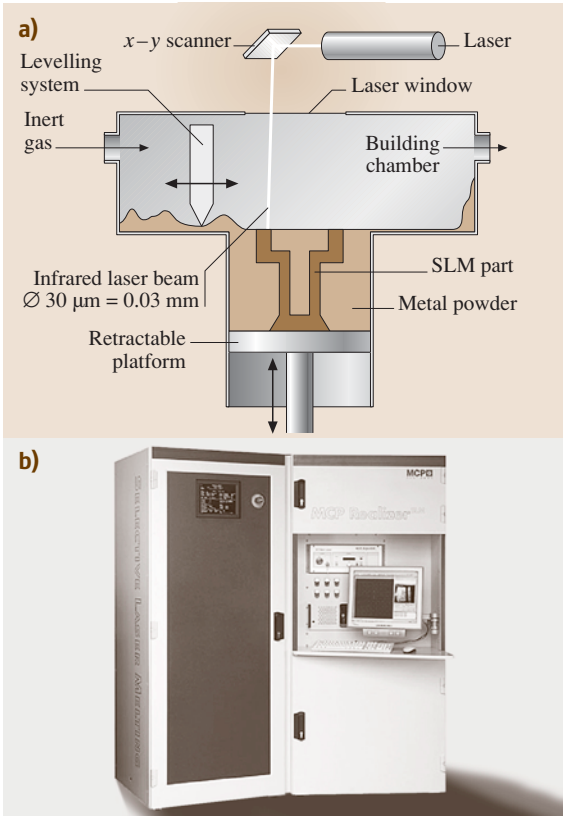


Fig. 7.336a,b SLM: (a) process principle, and (b) example of SLM machine – MCP Realizer II

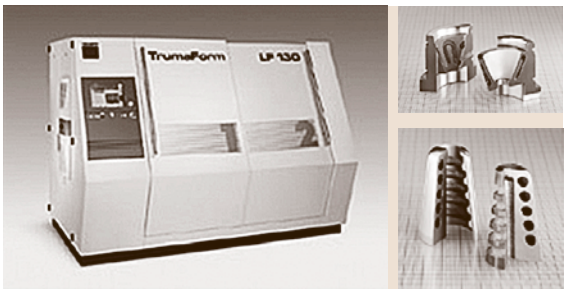


Fig. 7.337 TrumaForm LF 250 from Trumpf. Technical characteristics: full melting; layer thickness 50–200 µm; accuracy 0.1 mm; build volume ∅ × 130 × 160 mm or ∅ × 250 × 160 mm; speed 3–12 cm^3/h

Fig. 7.338 Phenix PM250 equipment

Fig. 7.339 M3 Linear from Concept Laser

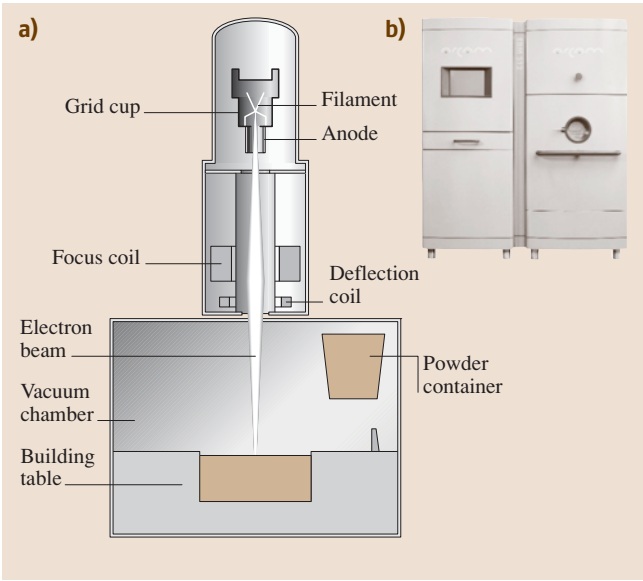


Fig. 7.340 (a) The EBM scheme and (b) the machine from Arcam

- Homogeneous components build up with up to 100% density.
- No necessity of postprocessing (heat treatment, infiltration).
- Fully automated process.
- Possibility of producing internal undercuts or conformal cooling channels.
- High process resolution and accuracy, low heat production.

Technical characteristics:

- Materals: stainless steel 1.4404 (316L), tool steel 1.2344 (H13), titanium (cp-titanium) and titanium alloys (TiAl6V4, TiAl6Nb7), aluminum (AlSi12, AlSi10 Mg), Cobalt-Chrome (CoCr ASTM F75)
- Full melting
- Layer thickness: 30, 50, 75, (optional 100) µm
- Accuracy (material depending): 0.1–0.2 mm
- Building volume: 250 × 250 × 250 mm
- Building speed (material depending): 5–7 cm^3/h

Direct Laser Forming Technology. In the direct laser forming (DLF) process metal powder is melted and homogeneous parts are produced without any binder. DLF allows for producing casting molds and tools that are impossible or very difficult to manufacture in the traditional way. This technology enables creating internal conformal cooling channels, offers the possibility of creating molds, saving materials and functional details.

The dual-chamber idea increases the TrumaForm's versatility [7.438]. During the cooling process in one chamber, the next job can be started in the second chamber with different powder.

Laser Sintering Phenix Systems Technology. The company PHENIX SYSTEMS has developed a rapid prototyping and manufacturing process using laser sintering that represents a true alternative at times when the performances of conventional manufacturing means can no longer meet the technical or economic requirements for product development [7.439].

Manufacturing volume is made up of a cylinder 250 mm in diameter and a height of 300 mm.

The standard materials used are metallic (stainless steel, tool steel, nickel, etc.) and ceramic powders (alumina, mullite, zirconia, etc.), which can be found on the powder metallurgy market.

Common characteristics of metallic parts:

- Accuracy of the produced parts: ±50 µm per 120 mm

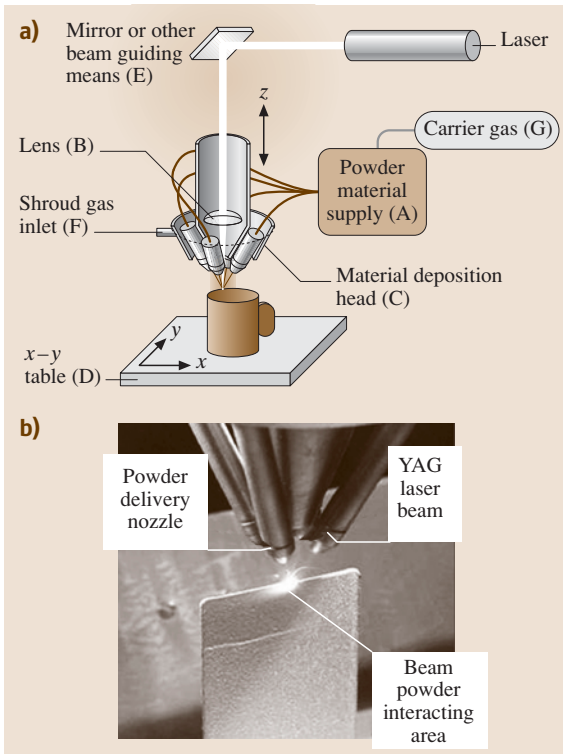


Fig. 7.341 (a) The LENS scheme and (b) realistic process

- Level of detail: 150 μm
- Postprocessing: polishing

Common characteristics of ceramic parts:

- Accuracy of the produced parts: ±50 μm per 120 mm
- Level of detail 300 μm
- Postprocessing: postsintering in the furnace

Technological characteristics:

- Fiber laser 50 or 100 W (high absorption of IR light)
- PM 100: diam. = 100 mm, H = 100 mm
- PM 250: diam. = 250 mm, H = 300 mm
- Building in a furnace: $T_{max} = 900\,°C$ with controlled atm.
- Building speed: 3–30 cm^3/h
- Layer thickness: 20–30 μm

Concept Laser. The M3 Linear is a modular system [7.440]. Apart from the additive manufacturing module it offers an erosion processing module and a marking module in one system.

Module 1 (laser cusing): the module for producing parts from metallic powders. It allows for building parts layer by layer from many materials (e.g., stainless steel and hot work steel). Metallic powder is melted to produce 100% component density. The exposure strategy allows for producing large-volume parts without deformations. A patented surface postprocessing ensures good surface quality and hardness.

Module 2 (3-D erosion module): the module for 3-D material erosion by a laser. It allows for erosion on free-form surfaces. The depth of the erosion process may be individually set by a laser-measuring sensor integrated with the machine software.

Module 3 (marking module): the module for creating signs on plastic or metal elements. It allows for laser marking and engraving on a wide variety of materials.

Technical characteristics:

- Materials: tool steel CL50WS I CL60DG, stainless steel CL20ES, titanium
- Full melting
- Layer thickness: 20 μm
- Accuracy: 0.1 mm
- Build volume: 250 × 250 × 170 mm
- Speed: 5 cm^3/h

Electron Beam Melting – EBM Technology. Arcam AB (Sweden) provides CAD to Metal® technology based on the electron beam melting (EBM) process originally developed at Chalmers University [7.441]. It is a powder-based method having a lot in common with selective laser sintering (SLS), but replaces the laser with a scanned 4 kW electron beam that produces fully dense parts. Materials available at present include H13 tool steel, Arcam low alloy steel, titanium alloy (Ti6Al4V), and pure titanium. Arcam low alloy steel is an easy-to-machine material for prototyping applications.

Parts are fabricated in a vacuum and at about 1000 °C to limit internal stresses and enhance material properties. The cooling process is also controlled to produce well-defined hardening. As with other processes, the parts require some final machining after fabrication, although the company indicates they feel their finishes might be somewhat better than those available from laser powder forming and other competitive processes. Arcam also says that processing in a vacuum provides a clean environment that improves metal characteristics. The EBM process may ultimately be applicable to a wider range of materials than competitive processes and also has the potential to offer much better energy efficiency.

Technical characteristics:

- Materials: steel A6 and H13, Ti6Al4V, Co-Cr
- Full melting
- Layer thickness: 50–200 μm
- Accuracy: 0.2–0.4 mm
- Build volume: 250 × 250 × 195 mm
- Speed: 10–60 cm^3/h

Laser Engineering Net Shape – LENS Technology. Laser engineered net shaping (LENS) technology developed by Sandia National Labs has been commercialized by Optomec [7.442].

This process is similar to other rapid prototyping technologies in its approach to fabricating a solid component by layer additive methods. However, the LENS technology is unique in that fully dense metal components are fabricated directly from raw materials, bypassing initial forming operations such as casting, forging, and rough machining. Parts have been fabricated from stainless steel alloys, nickel-based alloys, tool steel alloys, titanium alloys, and other specialty materials; as well as composite and functionally graded material deposition. Microscopy studies show the LENS parts to be fully dense with no compositional degradation. Mechanical testing reveals outstanding as-fabricated mechanical properties.

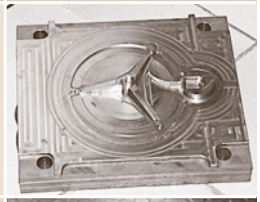
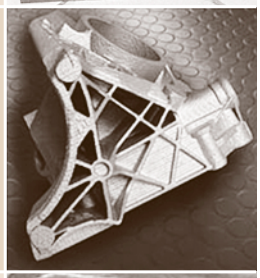
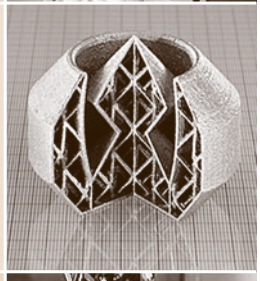
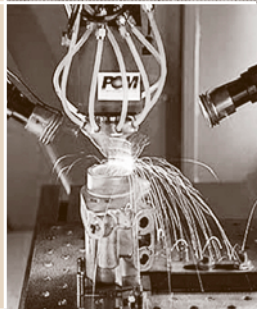
Technology	Company/model	Build technique	Materials	Technological characteristics	Samples
Layer machining stratoconception technology	Stratoconception/ STC 1510 www.stratoconcept.com	High speed micro-milling	Plastics, wood, PU … Aluminum	Manufacturing volume (mm): 1010 × 500 × Unlimited Accuracy (mm) 0.025	
Inkjet	ProMetal SR2 www.prometal.com	Three dimensional printing	Steel alloys, cooper alloys, tungsten alloys, nickel alloys	Manufacturing volume (mm): 1000 × 500 × 250 Printheads: 32 jets Building speed: 1100 cm^3 Layer thickness 120 – 170 µm Bronze infiltration: –> 40 % Bronze, 60 % Stainless steel	
Metal printing	SINTEF group / MPP www.mpp.no	Metal printing process	Steel, Iron, Ti, Ni, ceramic, … Tungsten, Molybden, …	Photo masking (light exposition based on 2 D section) Electrostatic attraction of the powder Electric discharge sintering or microwaves High speed (5 s / layer) Supports required No debinding or infiltration Density 92 %	
Ceramic laser-sintering	EOS / EOSINT S www.eos.info	Sintering	Foundry sand	Manufacturing volume (mm) 720 × 380 × 380 Layer thickness 200 µm	
Direct metal laser melting	Trumpf + EOS/ DMLM www.trumpf.com	Metal laser melting	Tool steel (1.2343), titanium (Ti-A16V4) and stainless steel (1.4404)	Full melting – full density Laser: 250 W / 500 W (Yb: YAG laser) Internal temperature: 500 °C inert atmosphere Automatic powder removal Layer thickness 50 µm Speed 5 cm^3/h Rough surface quality Build size: 250 × 250 mm	
Direct metal deposition	POM / DMD 505 www.pomgroup.com	Fused metal deposition	Nickel-based alloys such as inconel and hastalloy; cobalt, copper and tungsten-based alloys; cermets; stainless steels; tool steels and precipitation-hardened steels	Manufacturing volume (mm): 2000 × 1000 × 750	

Fig. 7.342 Specification of other RP technologies

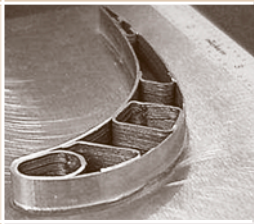
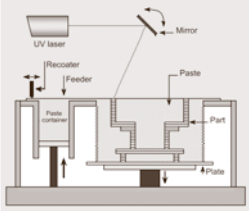
Technology	Company/model	Build technique	Materials	Technological characteristics	Samples
Metal deposition	H&R Technology/ PMD www.hrtechnology.com	Precision metal deposition	Titanium alloy Ti-6-2-4-2	No binder No infiltration No furnace step Large overhang surfaces without support structures Accuracy: 0.05 mm Speed: 7cm^3/h	
Laser sintering	3DSystems/ DSM Somos Technology/ Optoform www.3dsystems.com www.dsm.com	Laser sintering	Ceramics, composite tooling materials and thoughened plastics	Debinding step (polymer removing) Sintering step (densification & mechanical properties enhancements) Layers: 50 – 75 µm Filled polymer, metal & ceramic applications (RT&RM) Not applicable to conformal cooling channels: removal of paste into the channels is very complex	

Fig. 7.342 (cont.)

Technological characteristics:

- Materials: including steels, cobalt alloys, titanium alloys, nickel alloys, aluminum alloys, copper alloys, and other
- Near net shape (surface roughness)
- Functional graded materials
- Build volume: 1500 × 900 × 900 mm

Other RP Technologies

Apart from the technologies described above there is a multitude of methods and machines from other suppliers. Below, example technologies are classified with their industrial and research applications. More information on the world market may be found in reports published annually by Wohlers Associates.

There are more and more solutions and suppliers emerging apart from the technologies and machines shown above (Fig. 7.342). Fast progress in this area is to be expected because the issues of rapid manufacturing and customization are growing in importance. Better efficiency, accuracy, surface quality, and new materials will soon cause wider acceptance of these technologies and more industrial applications.

7.5.3 Reverse Engineering Technologies

Digitizing of 3-D Geometry

Digitizing means numerical notation, converting analog data into a digital form that can be saved in a computer's memory. The variability range of the analog data is subdivided into intervals (so-called quanta) and each interval is given a constant, averaged numerical value. The smaller interval, the higher the digitizing resolution and, at the same time, the higher memory consumption for storing the results. Spatial digitizing means numerical notation of a spatial, geometrical model shape in the form of coordinates of points located on its surface in a generalized coordinate system.

Two groups of methods of spatial digitization can be distinguished [7.443].

- *Contact* methods, which use a movable probe acquiring information in the form of coordinates of points on the examined surface, defined in a 3-D coordinate system. Data saving and processing are performed in a computer's memory. In these methods a coordinate measuring machine (CMM) can be used.
- *Noncontact* methods, where a physical model is scanned layer by layer, e.g., by means of a laser beam or X-rays. By converting the acquired contours and surfaces, a 3-D computer model is obtained, designed for further processing in a RE process.

In combination with the modern CAD systems and rapid prototyping technologies, digitizing has given a new meaning to such terms as *rapid product development* or *prototyping*, and in the complete cycle it can be called reverse engineering (RE) (Fig. 7.343) [7.444].

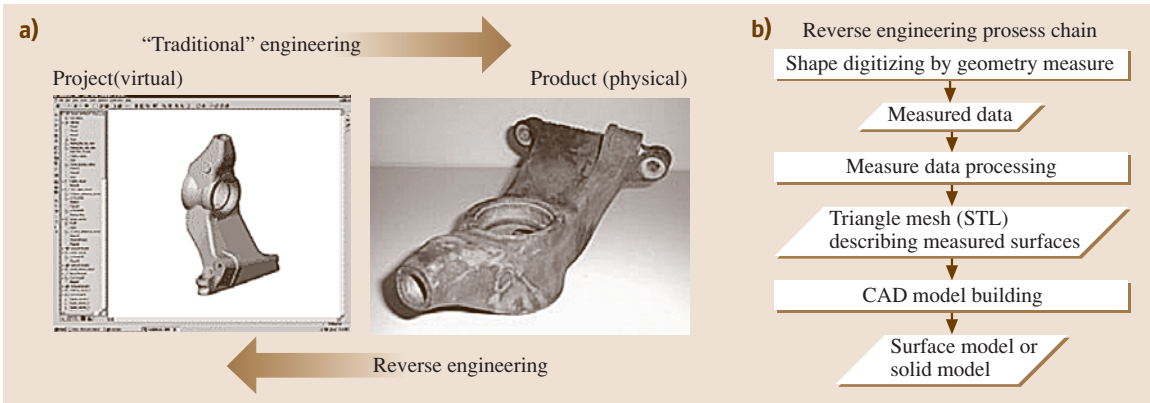


Fig. 7.343 (a) Reverse engineering idea and (b) basic operations at creating a 3-D CAD model

Basic digitizing methods and their application range with regard to their accuracy are shown in Fig. 7.344

Digitizing has created and is still creating new possibilities and its application range is still developing. Its dynamic development results simply from growing demand.

As a result of digitizing, spatial positions of a number of points located on the surface of an object are obtained, based on which a 3-D surface model is created in a CAD system environment [7.414].

The application software associated with machines for spatial digitizing allows exporting data in the most popular formats:

- DXF (points, polylines, polygon meshes)
- IBL (polylines)
- IGES (points, polylines, spline)
- OBJ (polygon meshes)
- STL (polygon meshes)
- VDA (polylines)
- XYZ (point coordinates)

Application of Reverse Engineering Techniques

The importance of numerical notation combined with the above-mentioned rapid prototyping techniques has immensely affected development of the RE that finds application in many fields, among others in:

- Reproducing technical documentation
- Building object models for repair or recovery
- Building a technological model on the ground of an industrial designer's pattern
- Analyzing competitive products in CAE systems
- Designing the shape of an object interfacing with existing objects

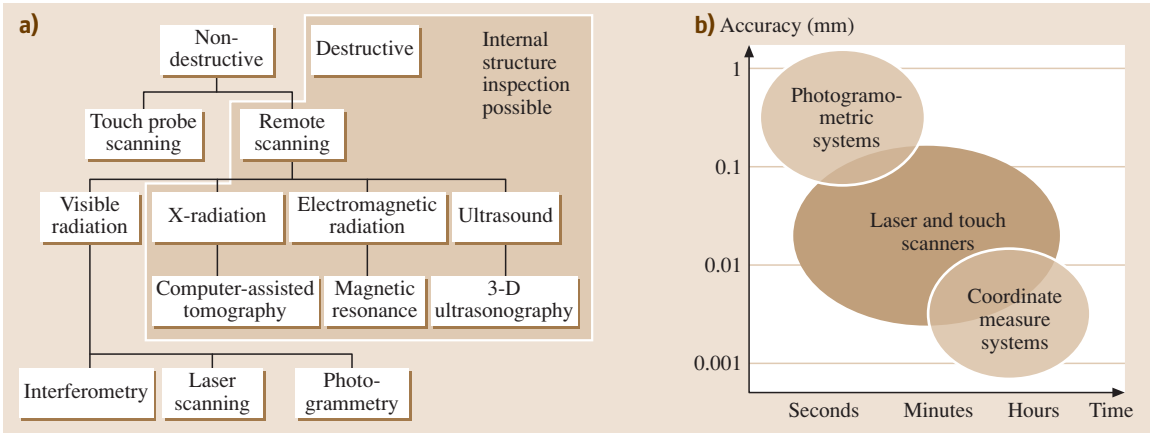


Fig. 7.344 (a) Methods of geometry digitizing; (b) exactness and time-consumption of digitizing methods (after [7.445])

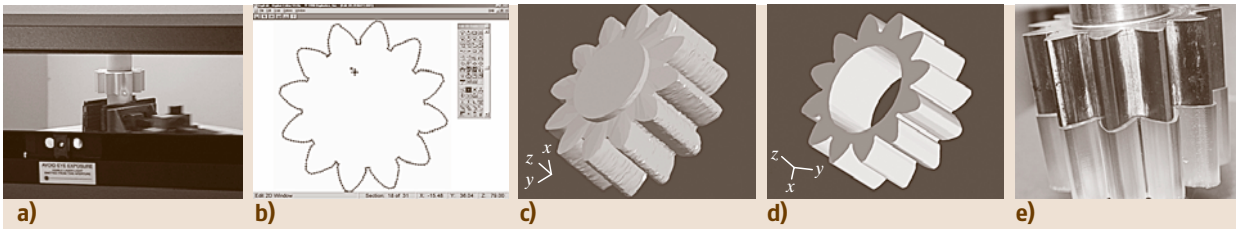


Fig. 7.345a–e Digitizing phases of a gear wheel: **(a)** placing on the scanner table and starting the digitizing process, **(b)** single contour obtained from the measurement, **(c)** geometrical STL model, **(d)** 3-D CAD model, **(e)** digitized gear and the gear made by SLA method after digitizing

- Updating design documentation after optimizing tests on a physical prototype
- Quality control – *first article inspection* aimed at evaluation of manufacturing processes
- Machine building, especially the automotive industry in the early development stages and in tool repair, quality inspection, reproduction of documentation, etc.
- Manufacture of packaging for new products of various, sometimes nontypical, shapes, designed by stylists
- Medal engraving and numismatics (shape reconstruction of coins and medals)
- Footwear and clothing industry (digitizing patterns and molds of footwear components, especially of customized products)
- Jewellery and souvenir industry (copying patterns and natural objects)
- Toy industry (making molds of the ground of artists' designs)q
- Art history (archiving and copying objects – sculptures, buildings etc.)
- New product development by means of rapid prototyping and rapid tooling technologies

A separate field of RE application is medicine, especially biomedical engineering. Owing to such equipment as a tomograph (computed tomography, CT) it is possible to record a 3-D geometry of a living organism or its internal organs [7.444, 446, 447]. Doctors and particularly surgeons have joined engineers who use the 3-D modeling. By means of models created on the ground of tomographic pictures they are able to prepare better for treatments or surgeries [7.414].

Here are some examples:

- Building a 3-D model from a series of flat sections of an object, obtained through computed tomography or magnetic resonance imaging (MRI) (Fig. 7.346)
- Planning complex surgical operations with the use of precise anatomic models built by rapid prototyping methods
- Designing an individual implant for a specific patient
- Reconstructing the geometry of biomedical objects for numerical analysis of their properties

Digitizing Techniques

Coordinate Measuring Machine (CMM). CMMs play an important role in precision measurements since, unlike in traditional measurements, they allow for

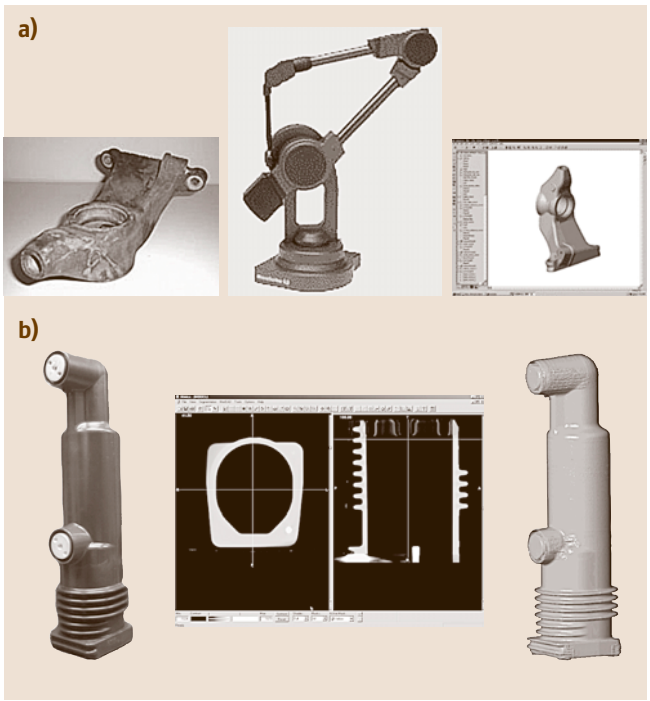


Fig. 7.346 **(a)** 3-D model of a digitized automotive part and **(b)** geometry of an insulator determined by computer tomography

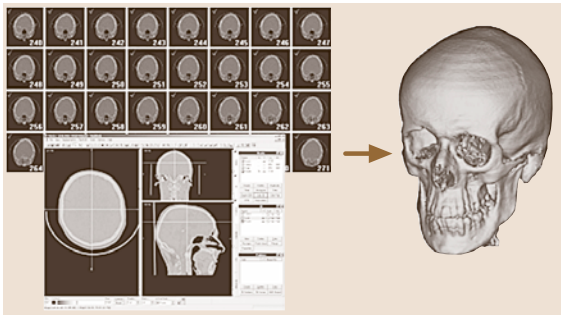


Fig. 7.347 Skull reconstruction on the ground of a set of CT pictures

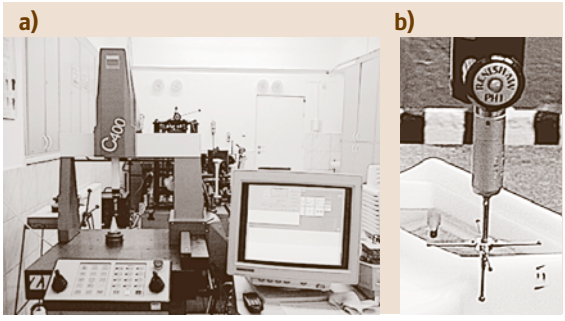


Fig. 7.348 **(a)** Coordinate measuring machine ZEISS C400 and **(b)** a measuring head, magnified

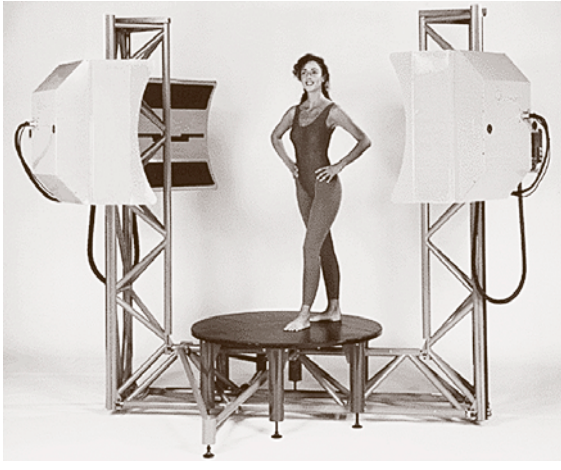


Fig. 7.349 Use of a three-dimensional scanner for clothing design (after [7.449])

quick, precise, and convenient, complex measurements of an object's geometry. Two basic methods of controlling those machines are available: manual (using a manipulator) and automatic (NC) supported by

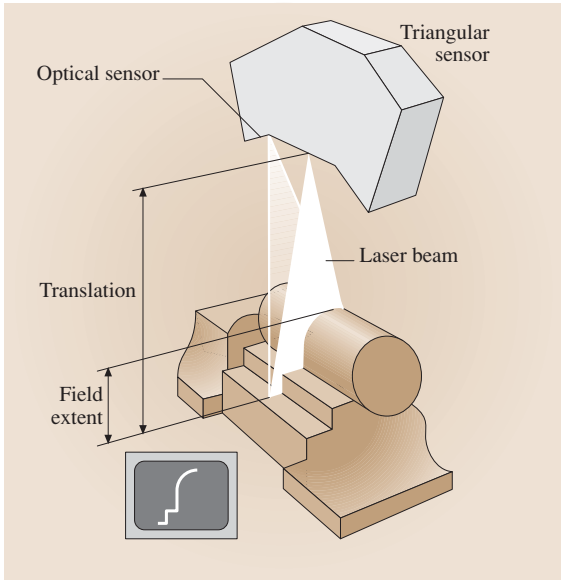


Fig. 7.350 Triangulation sensor used in laser scanners (after [7.450])

a computer program. The correct position of a measuring probe depends on the exactness of manufacture of the machine and its precise mechanisms [7.448]. Typical measurement accuracy is in the range of 0.005–0.0005 mm.

Digitizing of an object or its parts on a CMM is performed in the following way:

- The object is located on a table, the probe is selected, and the machine calibrated.
- The reference plane is determined.
- Characteristic points are indicated on the examined surface (using manipulators) as the basis for the measurement point mesh.
- The measurement step is defined, i. e., the distance between two consecutive measured points (the smaller the step the higher the accuracy)
- The manual or automatic digitizing procedure is started

Depending on the software, the results are processed by the machine controlling computer or exported to an external computer with suitable software, generally in IGES or XYZ formats.

The operation with the manual control is based on a *manually* indicated set of characteristic points of an object, i. e., the points that create simple geometric features adequate for reproducing the object by creating it in a CAD program in a 3-D space. This mode involves

an operator who sets the machine in motion by means of a control desk.

The operation in the NC mode does not require an operator after it is started. Such measurements are made by the machine executing a previously prepared program. The program includes data on the coordinate system, distribution of measuring points, sequence of measurement tasks, machine movement trajectories, calculations, and the form of result presentation. An NC program can be prepared in two ways:

- Teaching programming: the operator carries out all necessary measurements, considering the intermediate points that inform the machine of the idle movement trajectory. The machine saves the data and converts them in an NC programming language.
- OFFLINE programming: The program is created without the machine, in an NC programming language. The programmer inputs data and commands from the keyboard, in any text editor, and the instructions inform the machine of the courses and measuring tasks. The program can also be generated by a CMM module in a CAD system.

Laser Scanners. Three-dimensional scanners create computer models of existing objects. From among numerous developed variations of 3-D scanners and digitizers, the most popular use lasers [7.451]. They can automatically register shapes and colors of scanned surfaces with an accuracy reaching 0.125 mm, in a non-contact way, by means of an optical system consisting of a laser, movable mirror, and a sensor or video camera. The high quality, reliability, and operating speed of this type of equipment are reflected in their high prices.

Designers of 3-D laser scanners with video cameras use low-power lasers emitting orange-red or infrared light, revolving mirrors to deflect the laser beam and high-resolution CCD converters similar to those used in video cameras. The laser beam is deflected by means of a computer-controlled mirror. Image of the laser-illuminated place is recorded by a black-and-white or color video camera.

The scanning process runs automatically in one of three modes offered by the control program:

- Translation mode – the laser with the sensor moves along the arm, the model is fixed
- Rotation mode – the model rotates with the table, the laser head is fixed, and only the sensors move
- Adaptive mode – the laser head and the sensors move along the arm, the model is rotated to acquire the best *view* of the laser beam spot on the object's surface

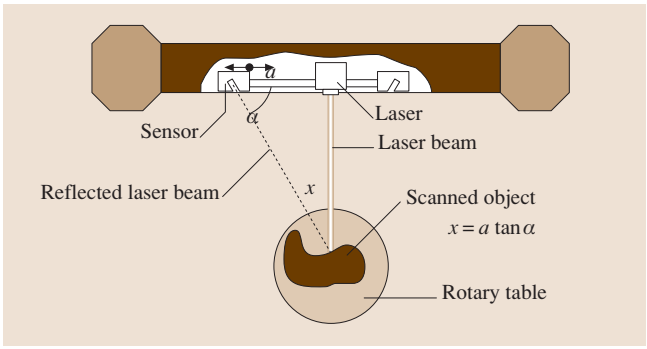


Fig. 7.351 Triangulation on the example of the Digibot II scanner

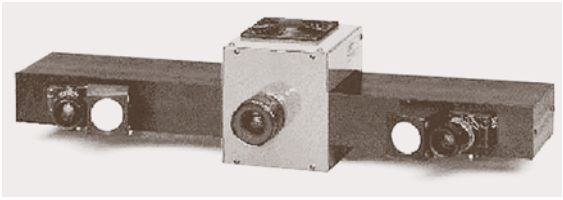


Fig. 7.352 White light scanner

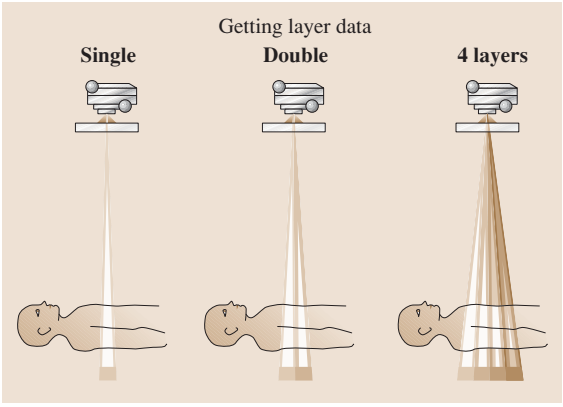


Fig. 7.353 Multilayer technology of computer tomography

In each of those modes the operator can set distances between individual layers (z axis) and numbers of points on each layer, determining the scanning resolution.

Scanners with White Light Triangulation (WLT). This method is an optical technique based on stereovision. In its function it is very close to the operation principle of laser scanners – the measurement data are calculated from the incident ray angle, intercepted reflected ray

angle, and locations of the light source and the recording camera. While laser scanners use a laser beam, WLT scanners use white light [7.451].

The main advantage of the process is that during the scanning process a mesh of measured points is recorded in the form of parallel layers. The laser scanners acquire the point data from one layer only. Image read-outs at various angles are combined and transformed into a 3-D cloud of points.

Computed Tomography. Computed tomography (CT) is a radiographic imaging method that uses a computer to reconstruct images from parallel cross-sections of an object. The 3-D object *image* is created from individual scans (slices of radiographic information) calculated for the analyzed object. The resulting image precisely represents the internal and external geometry of the object (Fig. 7.346). The distance between individual slices is selected by the machine operator depending on the object's degree of complexity. A single slice consists of a large number of points (pixels) in a 2-D image. Computer 3-D CAD models may be built by stacking the consecutive slices.

This method allows for measuring objects with complex shapes without physical access to their insides. It is a nondestructive method. Outside and inside dimensions can be measured with the accuracy of 0.002 mm. CT also delivers information on the density and composition of analyzed objects.

Apart from the original medical applications, CT is also playing an increasingly important role in industry. With equipment shown in Fig. 7.354 it is possible to transfer images of machine parts to virtual space with very high accuracy. A perfect example is the project conducted at the University of Munich [7.452].

The university undertook a project aimed at producing a casting mold for engine heads of an old car made in 1937 [7.453]. The first stage was to scan an existing, well-preserved head using CT. The 550 mm long

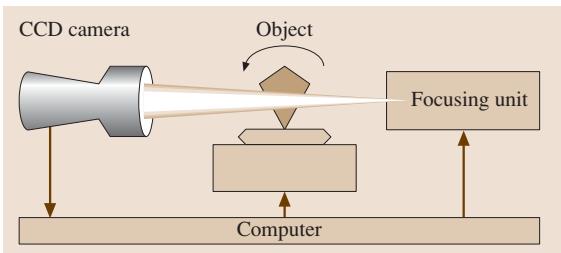
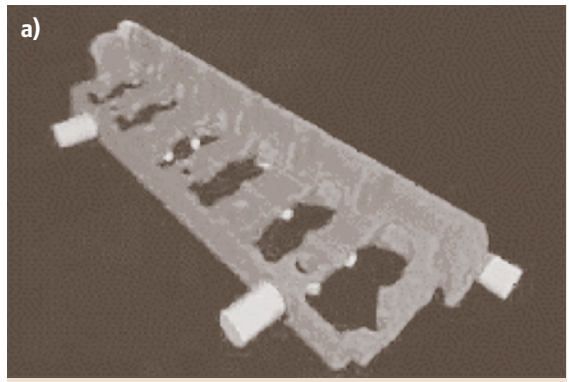


Fig. 7.354 Layout of a microtomograph used in industrial applications



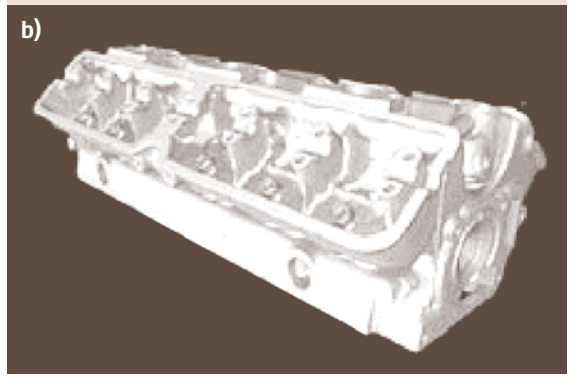


Fig. 7.355a,b Virtual 3-D models generated on the ground of CT scanning: **(a)** core and **(b)** head body

object was *sliced* into 1080 layers with a resolution of 1024 × 1024 each.

From those data the next step was to reconstruct the head body and with an array of internal passages. Both models were saved in the STL format; the first one was used for generating the mold model and the second for the core model. Both the mold and the core were made of sand in a laser sintering technique (SLS).

The finished mold was used to make a casting of an aluminum alloy. The cast body was subject to finishing on a five-axis milling machine. The final result is shown in Fig. 7.356.

Moiré Interferometry. Moiré interferometry (MI) is an optical method that has been used for many years in 3-D applications. It is a very exact tool (accuracy on the order of single micrometers) that determines surface profiles of objects as large as telescope lenses, car dashboards, or other difficult-to-measure components. This technique is based on illuminating the examined object with the Moiré light pattern and acquiring the reflected image by high-resolution CCD cameras. At

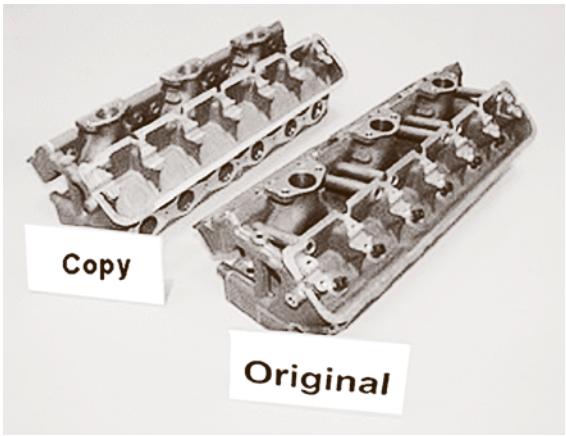


Fig. 7.356 Original head body and a copy based on CT pictures

present traditional MI has been developed into a scanning method.

Traditional MI consists of the following stages [7.454]:

- Illuminating the object surface with Moiré's pattern light
- Saving data from a CCD camera
- Shifting the pattern, very precisely, by a few micrometers
- Repeated saving data from the CCD camera
- Repeated shifting of the pattern by the same distance
- Finally, saving data from the CCD camera
- Reproducing the object topography with calculations based on the obtained picture

As a result of these steps each single point of the object surface has three pictures taken at three different light intensities. In this way, by simple computer calculations, an exact point location is obtained. The high costs and low capacity resulting from the long digitizing process (very precise shift of the pattern) did not allow this technique to be used on a large scale.

A solution is scanning Moiré interferometry (SMI), which has overcome the disadvantages of the traditional technique. The difference between the two concerns the procedure of acquiring points from the surface. Instead of a high-resolution 2-D CCD camera, the new process uses three-line scanning. Now the point-acquiring process is as follows:

- The CCD camera and the Moiré light pattern emitter are combined to create a special compact sensor.
- The object is moved in front or behind the sensor that reads the light intensity for each point.
- The object topography is created on the ground of calculated luminous intensity of each acquired point.
- The scanning process is continuous.

The following advantages were thus obtained:

- The linear CCD camera eliminates laborious picture taking for three subsequent positions.
- In combination with a movable manipulator, it is possible to process the acquired data continuously, at a speed equal to the recording speed on the video camera.

However, this has not changed the fact that this technique is useful for static objects only. Application of the SMI technique allows for digitizing objects with the same accuracy as the original MI but at a higher speed and with lower costs.

Reconstructions Based on a Set of Photographs (Photogrammetry). Photogrammetry is a technique of measuring 3-D objects from 2-D photograms. The term *photogram* covers both ordinary paper photos and images recorded with digital cameras or X-ray equipment. The results of photogrammetry may be:

- Models in the form of point coordinates
- Topographic and subject maps

Photogrammetry is a noncontact technique. It is popular in two application areas:

- Aerial photogrammetry used for building maps and digital land models
- Close-range photogrammetry used by architects (to check the condition of buildings and to determine their deformations and damage), archaeologists, and surgeons

In the photogrammetry technique three methods may be distinguished to obtain models and maps: from a single picture, from a series of pictures, and from stereophotogrammetry. All these methods involve using very expensive equipment like a metric or stereometric

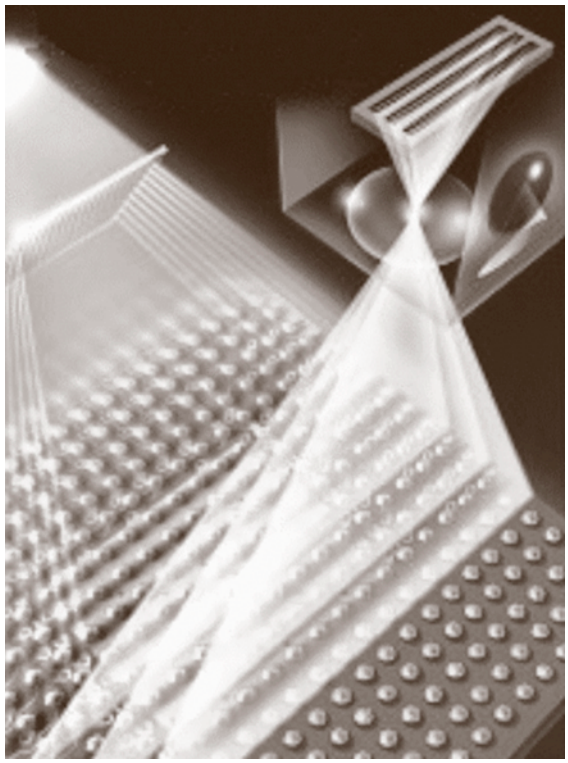


Fig. 7.357 SMI sensor in the three-line scanning process

camera. The acquired image is converted through complicated operations and computer calculations until the final results are obtained [7.455].

Black & White Scanning. Another digitizing method is used by Align Technology, USA. Since 1997 the company has been manufacturing orthodontic equipment in the system invented by them, Invisalign [7.456].

The manufacture of their teeth aligner starts from an impression of a patient's jaw, which is then used to make a plaster cast. Several such models are submerged into a block of a thermosetting plastic.

The hardened resin/plaster block is placed on a basic digitizing stage. On a device that could be called a *scano-milling machine* the block is repeatedly milled and then photographed. In each pass its height is reduced by 0.001 in. (Fig. 7.358).

A *stack* of black-and-white images of all layers is transferred to a computer, where, using a program developed by Align, a digital model of the jaw is built in 3-D space.

The computer models of patients' teeth are used by an orthodontist to generate series of aligners to be worn

Fig. 7.358 View at a single layer of a block

by patients to gradually correct the position of their teeth. The patterns for aligner manufacture are built in stereolithography.

7.5.4 Rapid Tooling Technologies

The pattern models and prototypes obtained with rapid prototyping methods are usually manufactured in small series for marketing and exhibition purposes or for experimental and service research. At this stage of product and manufacturing process development materials (or equivalents) and colors prescribed by the designer are used and the product is given suitable aesthetic features that should fully meet the features of series production. There are several commonly used rapid tooling (RT) technologies. They aim at providing tools (molds, dies) for manufacturing shorter or longer series of products in either specific processes or standard processes common in the production environment. The term rapid tooling covers various techniques of tool manufacturing, including forming inserts of injection molds from plastics, low-temperature melting metal alloys, or metallic powders.

Depending on the strength of the applied materials and their durability and application range, the subgroups of *rapid soft tools* and *rapid hard tools* can be distinguished among the tools manufactured by RT methods. The latter subgroup is characterized by higher durability and wider application range, and the properties close to those of molds manufactured by traditional machining technologies.

The techniques used in the manufacture of tools for mass production are based on highly efficient lost material machining called *high-speed cutting* (HSC),

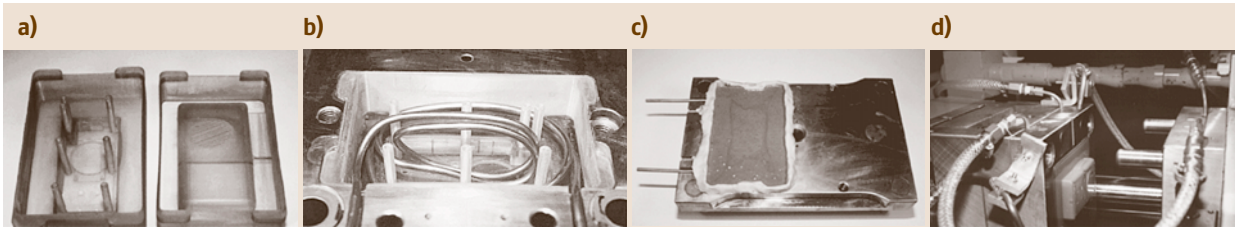


Fig. 7.359a–d Stages of creating an injection mold with cavities made by Direct AIM method: **(a)** stereolithographic forming inserts, **(b)** insert with cooling channels placed in the mold frame, **(c)** sealing the cooling channels in resin filler, **(d)** mold ready to be installed on an injection machine

high-speed milling, and *high-speed machining* (HSM). These technologies, whose development was prompted by new machine tools and machining strategies, allow for quick manufacture of dies and molds with complicated shapes from various materials, including hardened steels.

Direct Rapid Tooling Methods

Direct AIM. One of the direct manufacturing technologies of forming inserts is *direct AIM* (AIM = *ACES injection molding*), where forming elements are manufactured from epoxy resin hardened in a stereolithographic (SLA) process [7.412, 427].

The method was developed by 3-D Systems on the basis of model-building technology *accurate clear epoxy solid* (ACES). Injection molds made by this method are characterized by sufficient thermal and mechanical strength. The glassy temperature of such inserts is about 75 °C, which allows for obtaining 20 to 100 moldings of materials at forming temperatures of up to 300 °C. Here, the precondition is using a proper cooling system. An example of forming inserts made using *direct AIM* technology, including the mold-building stages, is shown in Fig. 7.359.

Direct Metal Laser Sintering. The direct metal laser sintering (DMLS) process, based on the SLS method, was developed by EOS GmbH and is marketed with the EOSINT M 250 machine (Fig. 7.360) [7.435]. This technology uses a mixture of bronze and nickel powders without additional bonding components. The powders have different melting temperatures, which fulfills the SLS principle based on the fact that the absorbed laser energy melts the low-melting phase (bronze) and wets the high-melting phase (nickel) that remains solid and develops as a result of an irreversible change in the crystal lattice. This ensures constant volume during sintering and thus results in high exactness of geometrical representation of the part. The occurring porosity within 25–45% is removed by infiltration of the mold with epoxy resin. The application range of this method can be expanded by the use of powdered steel materials. This group includes DirectSteel 50-V1, the material developed by EOS for the EOSINT M 250 Xtended machine. Steel powder with a grain size of 50 μm permits obtaining steel-forming inserts of injection molds for

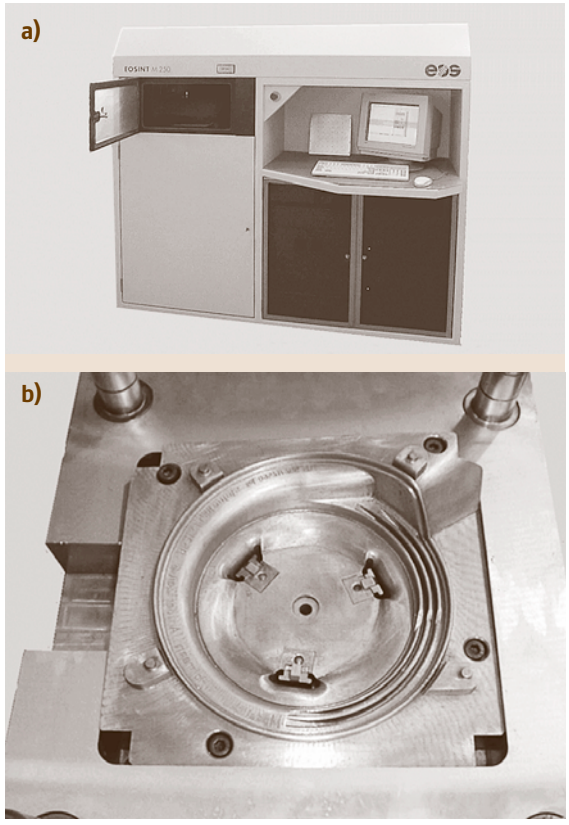


Fig. 7.360 **(a)** View of the equipment for laser sintering of metal powders EOSINT M 250 Xtended and **(b)** a prepared mold cavity

producing series of up to 100 000 pieces. This technique can also be applied for pressure die casting of highly abrasive plastics to obtain series of moldings of up to several hundred pieces.

Indirect Rapid Tooling Methods

Indirect RT methods use positive geometries of base models, created in rapid prototyping processes, to form *negative* mold cavities and dies. These technologies are applied in precision casting, vacuum molding, or in preparing mold cavities and dies by metal spraying.

The group of indirect RT technologies includes, among others [7.437]:

- Vacuum molding in silicone molds (*vacuum casting*)
- Building injection mold cavities by metal spraying (*metal spray*)
- Building molds of composite materials (*epoxy tooling*)

With regard to the different materials used in these technologies, they can be included in the *rapid soft tools* group in the case of vacuum molding and in the *rapid hard tools* group in the case of metal spraying and epoxy tooling. Vacuum molding is used for short prototype series, but metal spraying and epoxy tooling may be used for manufacturing longer series.

Vacuum Molding in Silicone Molds – Vacuum Casting. *Vacuum casting* (VC) is a technology that uses the environment and physical properties of a vacuum, both in the tool-preparing process and in the manufacture of prototype series [7.414, 437, 457]. The mold is made from two-component silicone resins that fill the space around a pattern model in the molding box. The entire tool-forming process runs in a vacuum, securing very accurate representations of the macro- and microstructure geometry of the pattern model. During the process, maximum negative pressure in the vacuum chamber reaches 0.5 mbar. All manufacturing stages of this technology are shown in Fig. 7.361.

The cycle, beginning from pouring into a mold, is repeated for each piece of the product manufactured. This technique allows for manufacturing of 5 to 30 pieces of a prototype product. The number depends on the product's degree of complexity. With very geometrically complex, thin-walled products, this number will always be smaller. The basic advantage of this tech-

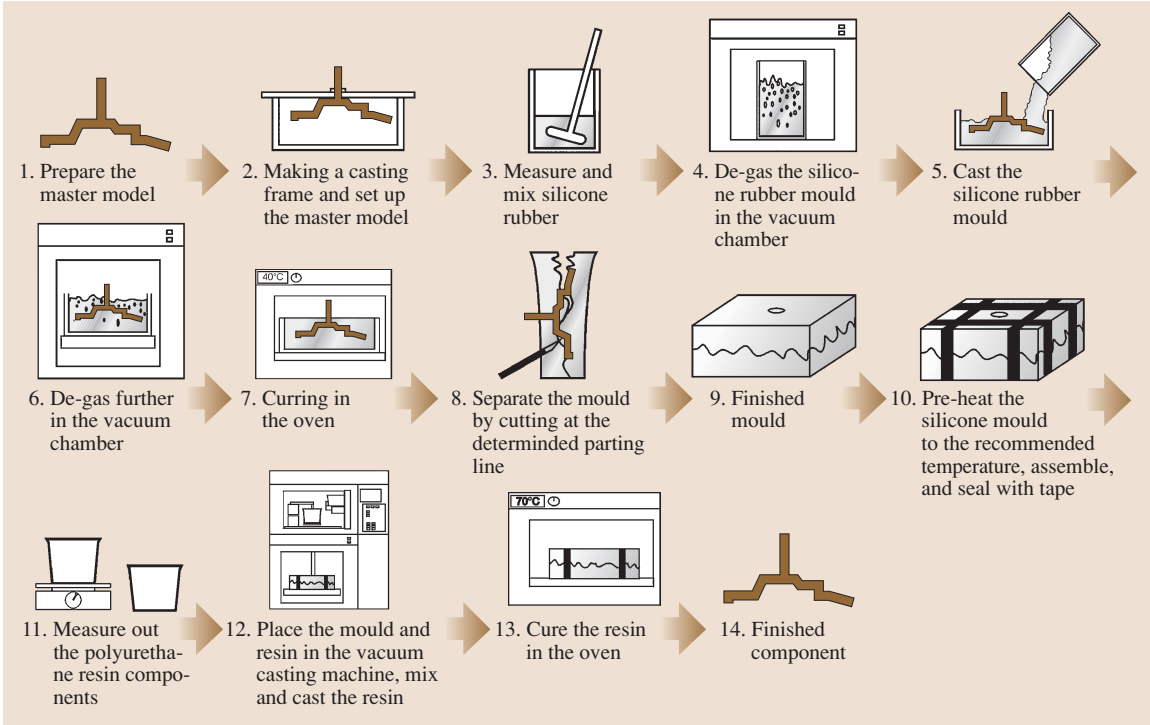


Fig. 7.361 Stages of vacuum molding in silicone molds

Fig. 7.362a–c Examples of vacuum casting application: **(a)** prototype products, **(b)** equipment for vacuum molding and **(c)** silicone molds

nology is its low cost, being about 3–6% of the cost of traditional technologies used in injection molding, and short production cycle counted in hours instead of months (about 3% of the time required in traditional technologies).

The cavities of silicone molds are used for gravity casting of prototype series products. The materials used for prototypes in this technology are two-component epoxy and polyurethane resins. They allow for obtaining prototype models with mechanical properties close to those of parts manufactured with injection molding. The resins used in VC imitate properties of typical thermoplastics used in plastics processing.

The equipment used in VC technology is shown in Fig. 7.363.

Characteristic features of silicone resins for molds in VC technology are presented in Table 7.57 [7.437, 458]

These materials, with their properties close to the thermoplastics family used in plastics processing, e.g., ABS, polyethylene, or polyvinyl acetal, allow for manufacturing a variety of prototype parts with a wide range of design and mechanical properties at full color scale [7.437].

Fig. 7.363 Equipment for vacuum molding 001ST made by MCP-HEK

With regard to the reaction of the mold material with the cast resins resulting in surface brittleness of mold cavities, the durability of the molds is limited to 20–30 cycles; however, the series repeatability is guaranteed for the first 10 pieces.

Figure 7.364 shows an example of an SL model with VC made in a silicone mold.

Metal Spray Technology. Another representative of the indirect *rapid tooling* techniques, of the *rapid hard tools* group, is *metal spray* technology. It allows one to prepare injection mold cavities, dies, and stamps for plastic forming of sheet metal [7.437].

Table 7.57 Parameters of two-component resins used in tool manufacture by VC processes

Component Property	ESSIL 241	ESSIL 244	VTV750/ CAT750	VTV800/ CAT800
Manufacturer	AXSON	AXSON	MCP-HEK	MCP-HEK
Setting time at 25 °C (min)	90–120	ca. 105	100	120
Density after hardening	1.10–1.14	1.10	1.2	1.2
Shore hardness	40 A	40 A	40 A	38 A
Ultimate elongation (%)	300	–	400	380
Linear shrinkage (%)	< 0.1	< 0.1	< 0.1	< 0.1

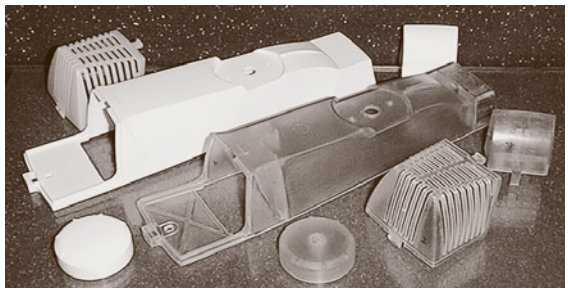


Fig. 7.364 Element of a prototype series with the original SL prototypes

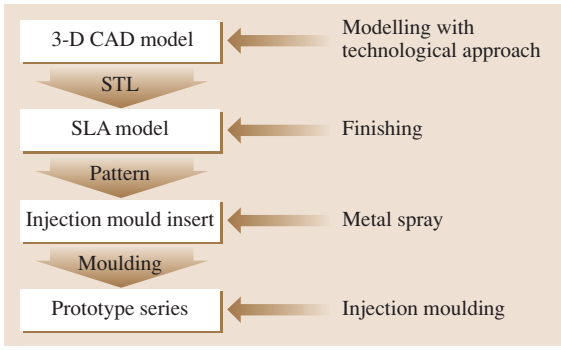


Fig. 7.365 Diagram of the process of prototype series manufacture by means of an injection mold made by metal spraying

The main advantage of this technology is that it allows for very quick preparation of mold inserts and dies based on a shell mold obtained by metal spraying of the pattern model previously prepared by, e.g., RP methods. The mold prepared by spraying the pattern model fully represents all the details of the model, with respect to both geometry and surface structure. The individual stages of this technology for injection mold building are shown in Fig. 7.366. Tools for mechanical working of sheets are manufactured in a similar way. The durability of a mold is estimated for 2000 to 10 000 pieces of finished product, depending on their geometrical complexity, applied materials, and process parameters. The typical time needed for mold preparation is 1 to 2 d, and costs do not exceed 15% of those estimated for mold manufacture by machining or electromachining.

When this method is used to manufacture dies and stamps for mechanical working of sheets, the processing sequences are similar, except that the process should consider the properties and thickness of the sheet to be processed [7.437].

Application of this technology is especially effective in both the initial and final stages of product prototyping, even including cars.

The range of materials that can be used with this method includes low-melting alloys of zinc, copper,

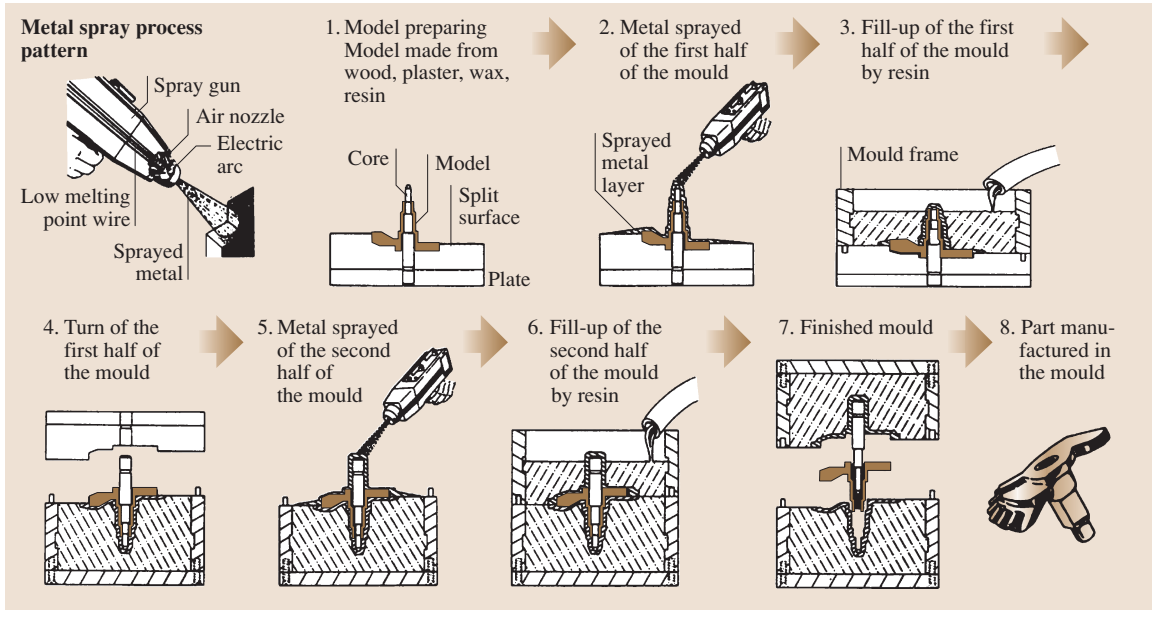


Fig. 7.366 Application of RT technology for preparing an injection mold cavity

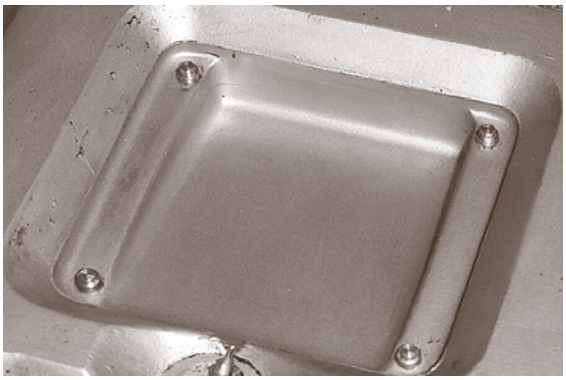


Fig. 7.367 mold cavity made by metal spray technique with brass inserts (after [7.414])

and aluminum, as well as steel. In the case of steel, it is necessary to prepare ceramic heat-resisting pattern models, which makes the tool-creation process more complicated. However, this method is applied in cases requiring higher heat resistance and mold durability. These features represent an advantage in tool preparation in comparison to the nonferrous metals molds.

An example of a mold cavity with brass inserts is shown in Fig. 7.367.

The metal spray technology based on low-melting alloys of tin and zinc was developed by the MCP-TAFA company. The method of building shell cavities of injection molds, stamping dies, and stamps for sheet processing was developed on the basis of the arc spraying machine TAFA-8830.

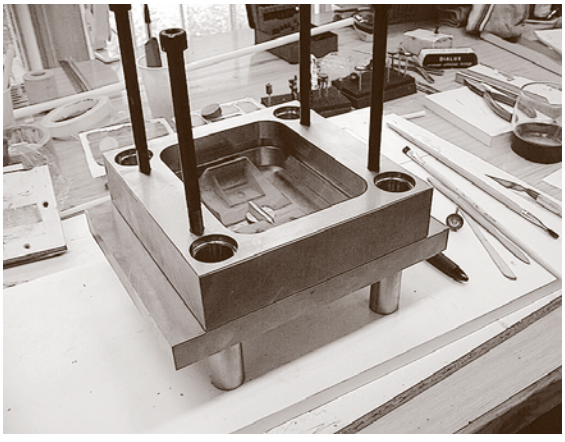


Fig. 7.368 View of a model with shaped parting plane, fixed in the mold frame

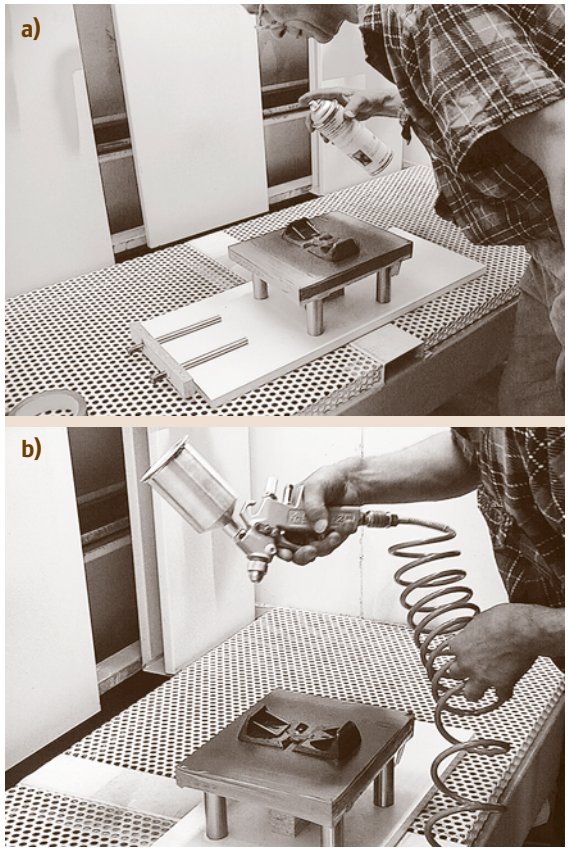


Fig. 7.369 **(a)** Stages of applying layers of graphite and **(b)** glassy separator

Apart from low-melting alloys developed for RT technologies, the spray gun TAFA 8830 also permits spraying such materials as copper alloys or steel. The process of pattern model preparation to be covered with an alloy layer includes the following stages:

- Giving the model the required properties of surface micro- and macrostructure, like roughness and surface quality.
- Fixing and shaping the parting planes.
- Covering the model with separators to enable dismembering the mold and removing the pattern model. Models with shaped parting planes are first covered with a graphite-based agent and next with a thin layer of another separator.

The spraying process (Fig. 7.370) should be continued until the shell is 3 to 5 mm thick.

Materials used in this method include, among others, low-melting tin and zinc alloys like MCP 200,

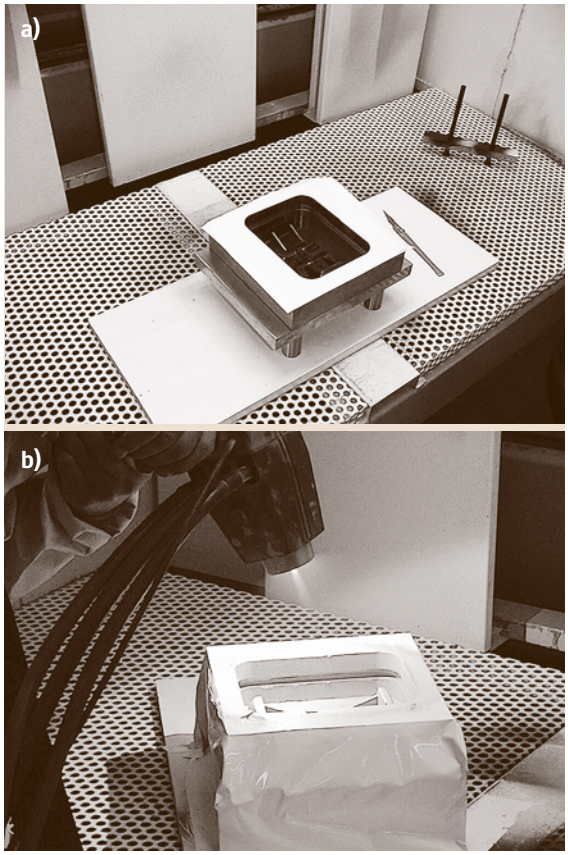


Fig. 7.370 (a) Metal spray of mold cavities prepared for spraying. (b) Course of the spraying process

TAFA 204M (MCP 400), and TAFA205. A view of a model with a shaped parting plane, covered with a separator and ready for spraying the cavity of the first half of the mold, is shown in Fig. 7.371.

The Costs of molds manufactured by RT techniques are much lower and, in the case of the technology developed by MCP-TAFA, comprise ca. 23% of the costs of molds prepared by traditional machining. Owing to this difference and to the time of mold preparation, the RT techniques are an effective tool in the initial stages of product development, significantly reducing the cost and time of new product implementation.

Epoxy Tooling, Composite Tooling. Another one of the *rapid hard tooling* technologies is a method of forming injection mold cavities, based on applying composites of epoxy resins and metal powders, usually aluminum [7.437]. This method permits reducing by ca. eight- to tenfold the time required to prepare process

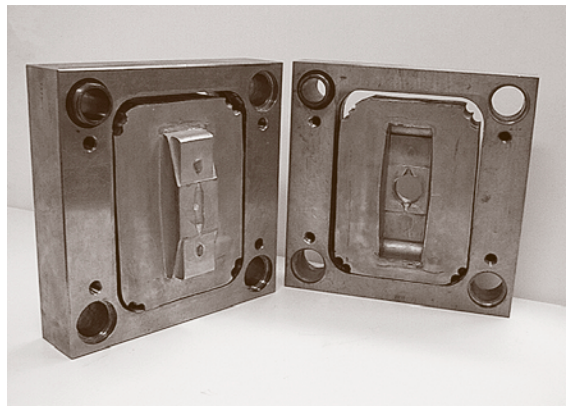


Fig. 7.371 Shell injection mold with cavities made by metal spraying (after [7.414])

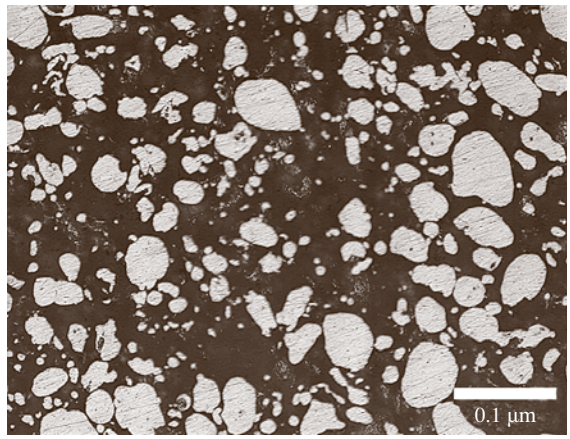


Fig. 7.372 Microstructure of epoxy resin reinforced with aluminum powder. Volume fraction of powder is 34%, grain size from 2 µm to 0.7 mm. Light microscope (after [7.414])

instrumentation for injection molding of plastics. The mold cavity preparation cycle includes:

- Preparing a model (e.g., RP technology) and parting planes
- Pouring the first half of the mold and sealing the cooling ducts
- Preparing the second half of the mold
- Pouring the second half of the mold with cooling ducts Fig. 7.373 [7.414]

With regard to the process of mold cavity preparation that is carried out in a vacuum, cavities prepared in this way can faithfully reflect even complicated shapes (with a large number of roundings and freely shaped

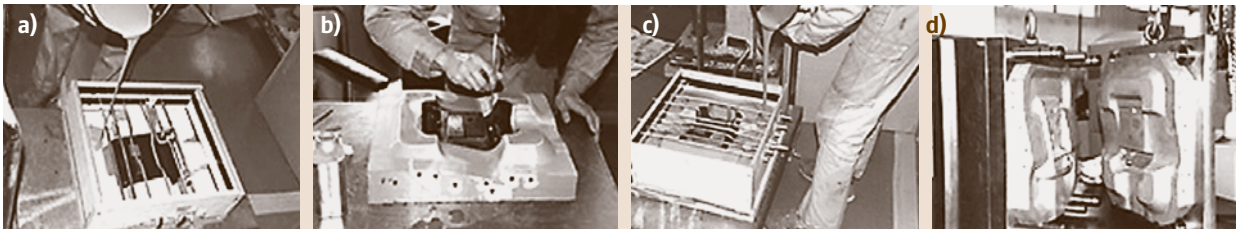


Fig. 7.373a–d Phases of building a composite injection mold in the rapid tooling technology: **(a,c)** filling the space above the model and parting plane with epoxy composite with aluminum filler, **(a)** finishing the prepared first half of the mold cavity, **(d)** finished mold ready to be installed on an injection molding machine

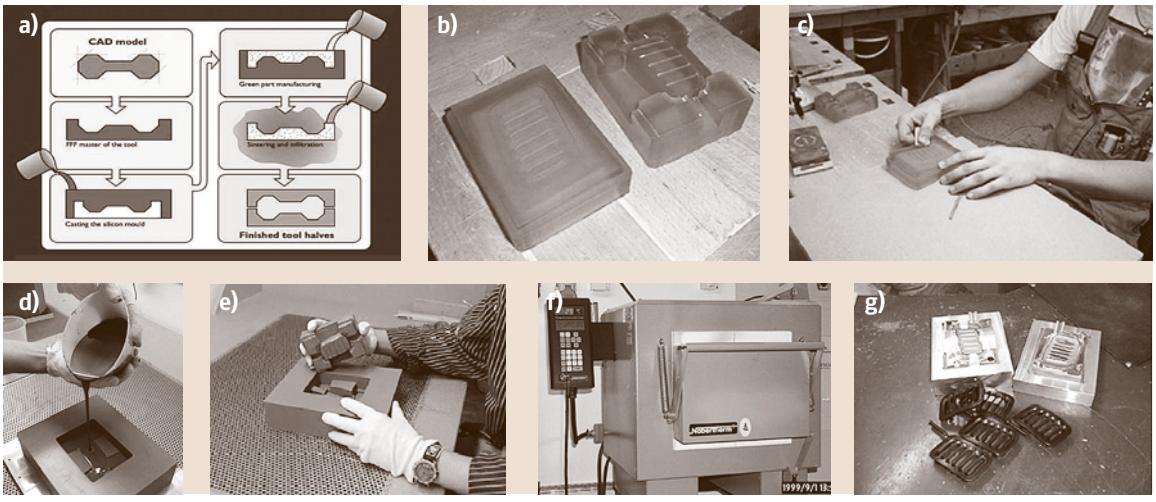


Fig. 7.374a–g Stages of the MetalCopy process

surfaces) that, in the case of traditional methods, would require multiaxial milling. Optionally shaped cooling ducts permit optimum heat abstraction from the cavity area filled with plastic in the process of injection molding.

METALCOPY Technology (IVF Sweden). A Swedish research institute and a Swedish SME have developed a new RT and manufacturing technology called MetalCopy [7.459]. MetalCopy is especially suited for demanding tool inserts and complex geometries and it dramatically reduces costs for producing tools/inserts and cuts lead times.

MetalCopy is a cost-efficient and RT manufacturing technology for injection molding dies for short series production. It is the ideal solution, particularly for applications where traditional methods such as milling, EDM (engineering development model), laser, and epoxy fail. MetalCopy produces durable injection molding tools and inserts for small series, bridge tools, and preseries production, up to 100 000 parts. The method is particularly suitable for complex insert geometries with deep holes and slots.

MetalCopy's rapid tooling process involves the following steps:

1. A primary master of the tool is produced by an accurate rapid prototyping machine.
2. A silicone negative is cast from the rapid-prototyped positive master.
3. The silicone negative is used to produce a *green* secondary master in a mixture of steel powder and binder.
4. The green part is sintered and the pores between the metal powder grains are filled by a low-melting-point metal.

5. The die tool is finished by traditional machining methods, such as drilling for ejector pins, fitting the tool halves together, etc.

The method's strength lies in its ability to deal with the most demanding requirements, such as complicated parts that cannot be milled, serial volumes far above what can be achieved by epoxy tools, or where an insert with good heat conductivity is needed. Main advantages: MetalCopy is particularly suitable for multi-insert tools when the insert geometry is identical. The geometry of this insert is characterized by a great number of thin and deep grooves and slots. This is a good example of where MetalCopy is preferable for manufacturing.

Specification:

- Production: short series and up to 100 000 parts
- Accuracy: 0.1–0.5 mm (0.2%)
- Surface roughness: $R_a < 2\,\mu m$
- Good thermal conductivity
- Cost efficient, especially for complex geometries

More information about these technologies may be found in reports published annually by Wohlers Associates, proceedings of Euro-uRapid conferences, and producer Web sites.

7.6 Precision Machinery Using MEMS Technology

MEMS (microelectromechanical systems) is one of the new approaches to producing or assembling precision machinery with electronics. Originally, MEMS emerged from the silicon process, which is a combination of CVD (chemical vapor deposition), RIE (reactive ion etching), photo or electron beam lithography systems etc.; it is a concept to create small mixtures of machines and electronics regardless of fabrication technology.

7.6.1 Electrostatic-Driven Optical Display Device

Various types of optical switching devices have been proposed and developed. Most of them are intended for switchboard purposes – typically optical-fiber communiation [7.460]. For display purposes, the digital micromirror device developed by TI [7.461] is a good example of a commercial application of MEMS. In the device, an application of mechanical action to the display purpose is one of the promising potential capabilities of MEMS. In particular, the frequency response of the mechanisms reaches more than hundreds of kilohertz because of their small mass.

Two examples of optical switch are shown in order. An interferometric display device (IDD) is composed of a small set of interferometers to control contrast and color [7.462, 463]. An evanescent coupling display device (ECDD) is an application of near-field optics for switching light perpendicularly. The principle of the device, its manufacturing process, and experimental results are shown.

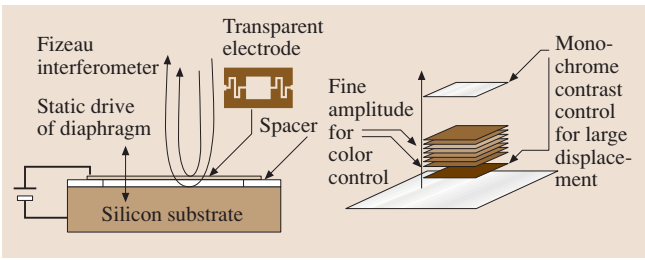


Fig. 7.375 Principle of the interferometric display device

Principle of the Interferometric Display Device

Figure 7.375 shows the schematic construction of the IDD. An SiO_2 half-mirror suspended by leaf springs forms a capacitor to the base plate with a gap of $(n + 1/4)\lambda$ to construct a Fizeau interferometer. When a DC voltage is applied to the capacitor, the half-mirror is attracted to the base plate balancing the leaf springs, and the reflection beam from the base plate interferes with the beam reflected from the half-mirror. When a monochromatic light source illuminates the device, the contrast is changed periodically according to the displacement of every $\lambda/4$. In the case of a white light source, color variation is realized with a fine half-mirror positioning control. Integrating this device like a dot-matrix, a reflection-type display can be constructed. The device has the potential to triple the pixel density compared to RGB space separation displays. Also, static driving consumes little energy, and the mirrors has fast response because of their light mass of tens of micrograms.

7.6.2 Design of the Device

The displacement of the mirror z is a balance of the leaf-spring force and the electrostatic force F. It can be expressed as

$$F = \frac{1}{2}\epsilon S\left(\frac{V}{d-z}\right)^2 = kz\,, \tag{7.212}$$

where ϵ is the dielectric constant, S is the area of the half-mirror, V is the applied voltage, d is the gap length without voltage, and k is the spring constant of the leaf spring. The resonant frequency of the mirror f_n is calculated by

$$f_n = \frac{1}{2\pi}\sqrt{\frac{ml^3}{192EI}}\,, \tag{7.213}$$

where m is the mass of the leaf spring, l is the length of the leaf spring, E is the Young modulus of the spring material, and I is the moment of inertia of the leaf spring. This predicts that the relationship between voltage and displacement is nonlinear.

The maximum contrast of the device is obtained when the intensity of the beam reflected by the half-mirror and that reflected from the bottom is equal. Suppose the following: half-mirror reflectance R_h, transmittance $T_\mathrm{h}(\simeq 1 - R_\mathrm{h})$, and reflectance of the bottom mirror R_b; the relationship is expressed as

$$R_\mathrm{h} = T_\mathrm{h}^2 R_\mathrm{b}\,. \tag{7.214}$$

This is the optical requirement for the device to get the maximum contrast.

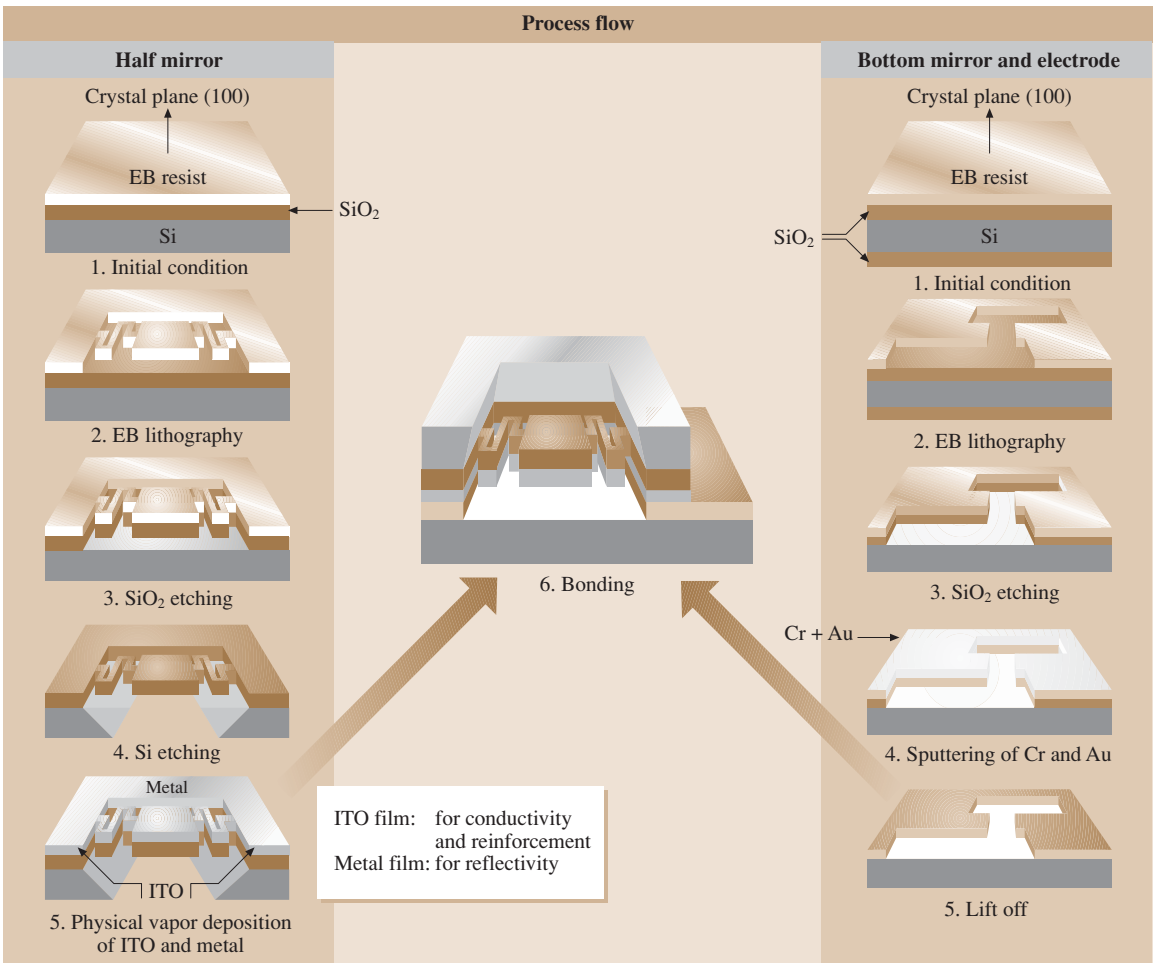


Fig. 7.376 Fabrication process of the IDD

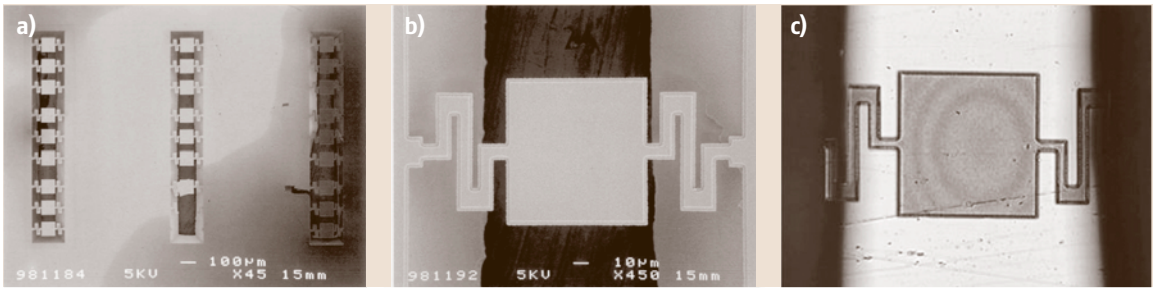


Fig. 7.377a–c Example of the (a) IDD array, (b) IDD pixel and (c) interferometric fringes

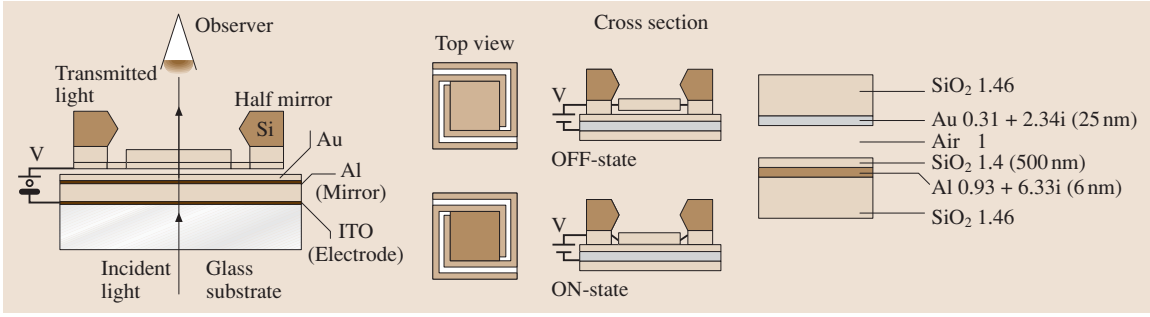


Fig. 7.378 Improved construction of the IDD based on Fabry–Perót interferometer

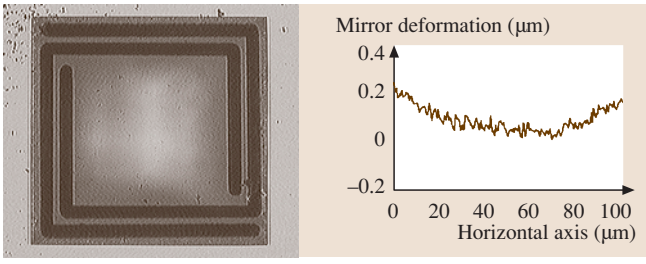


Fig. 7.379 A spiral leaf-spring structure and cross sectional deformation

Fig. 7.380 RGB pixel image

Fabrication of the Device

The half-mirror is the key element of this device and it should have a fine parallel movement to the base plate. A double-leaf-spring-suspended mirror is used for this purpose. A 100 μm square mirror is fabricated from a thermally oxidized silicon film of 1 μm thick, with a spring length of 400 μm, which is folded for a wider pixel aperture. A theoretical resonant frequency of 27 kHz is calculated for the film thickness, a beam width of 5 μm, and an applied voltage of 45 V. Thus, a displacement of the mirror more than a quarter of the wavelength can be obtained by the application of low driving voltage to the mirror gap.

Figure 7.376 is the process flow for manufacturing the IDD. A wafer bonding technique has been employed to adjust the mirror gap easily. The half-mirror and the bottom mirror electrode are fabricated separately. Then the gap is controlled by the film thickness of the SiO_2 as the spacer.

Performance of Prototype IDD

Figure 7.377 is a fabricated example of the IDD. For this prototype, a half-mirror dimension of 100 μm square, a leaf spring width of 10 μm, and a mirror gap of 5 μm are realized. A driving DC voltage of 15 V and a resonant frequency of 5.3 kHz have been obtained. When the device is illuminated by a halogen lamp, interferometric fringes are observed on the half-mirror, instead of uniform contrast change. This originates from the deformation of the half-mirror, because of its thin thickness and short leaf-spring length compared to the width,

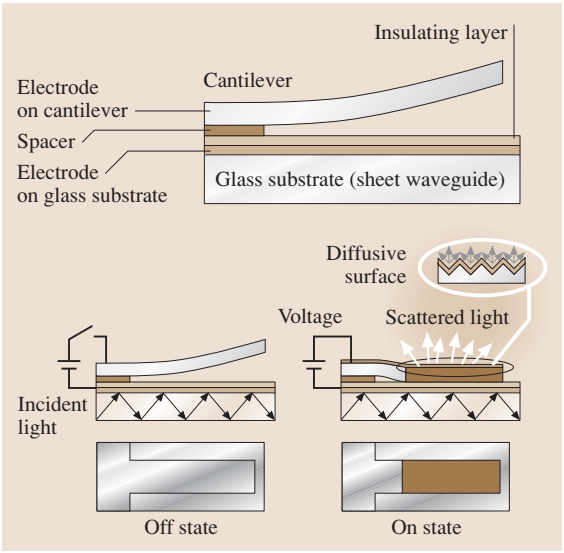


Fig. 7.381 Principle of the evanescent coupling switching device

and should be improved by film thickness increase as well as structural modification.

Improvement of the IDD

The principle of the device is alternated from Fizeau interferometer to Fabry–Perót interferometer, shown in Fig. 7.378, for easier optical adjustment for high contrast and illumination. By an experimental optimization, a combination of gold and aluminum film in the figure was found to be the best for satisfying (7.214). Also, to achieve smaller deformation of the half-mirror, a spiral structure, shown in Fig. 7.379, is adopted to realize longer leaf-spring length so that the deformation may be concentrated only on the leaf spring. With a twofold film thickness of 2 μm surrounding the half-mirror, the deformation of the mirror itself was improved to 0.2 μm, which is comparable to a quarter of the wavelength.

By a CCD video capture system, the RGB contrast of the IDD is evaluated. In Fig. 7.380, RGB contrast is obtained for the halogen illumination. Although the pixel image still has a distribution in both color and contrast, no interferometric fringe is observed for this construction thanks to smaller film deformation.

The average contrast of a pixel for each color $C_{R,G,B}$ is calculated as

$$C_{R,G,B} = \frac{I_{max} - I_{min}}{I_{max} + I_{min}} , \qquad (7.215)$$

where I_{max} is the maximum and I_{min} is the minimum intensity of the pixel. The obtained contrasts for the figure were $C_R = 0.18$, $C_G = 0.03$, and $C_B = 0.08$. The contrast dependence on color may be attributed to the interferometric layer, which is designed to be fit for only one wavelength.

Frequency response was examined by a function generator and high-voltage power source. Up to 100 Hz, the device follows the input voltage, and for higher driving frequency, it has a phase delay because of the squeeze damping effect of the air. For a driving fre-

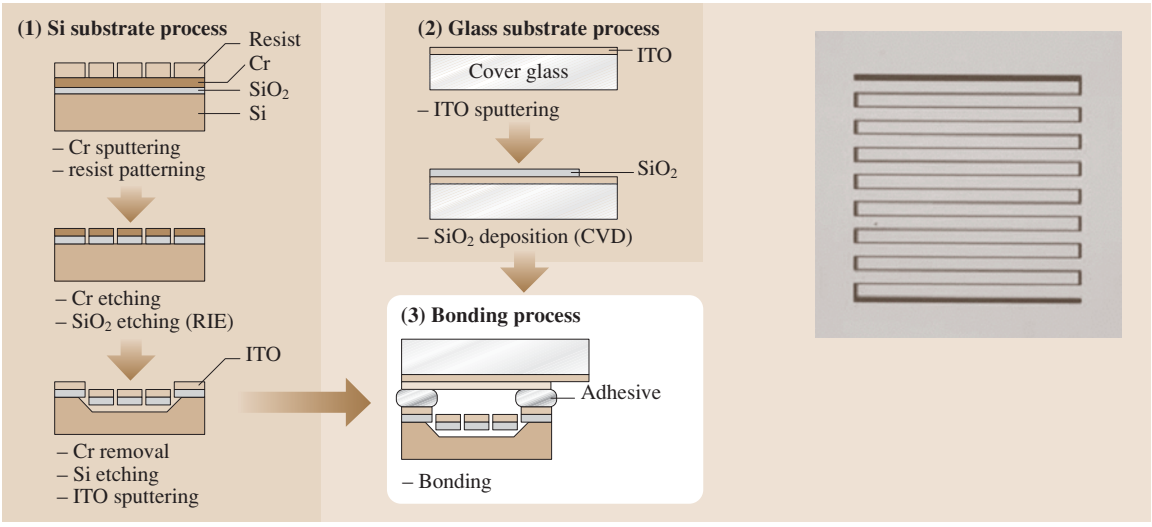


Fig. 7.382 Fabrication process of the ECDD

quency of 2 Hz and a voltage of 10 V, a maximum contrast of 0.3 was obtained.

7.6.3 Evanescent Coupling Switching Device

Fabrication of the device is very difficult because of process problems such as deformation of the half-mirror, adhesion of the mirror to the substrate, etc. An alternative is the evanescent coupling display device (ECDD), which uses microfabricated cantilevers for optical switching and free from mirror deformation [7.464–466]. The principle of the device, its fabrication process, and evaluation are presented in this section.

Principle of the Evanescent Coupling Display Device

The ECDD is fabricated on a glass substrate together with cantilevers as micro-optical switches. The substrate woks as a sheet waveguide propagating light horizontally, and the cantilevers are driven electrostatically by DC voltage to engage in or disengage from switching action. When the cantilevers are attracted to the substrate making mechanical contact, light energy in the sheet waveguide transfers to the optical cantilevers by the evanescent coupling, even if there is a small gap within the wavelength. The transferred light also propagates in the cantilevers horizontally; however, it is scattered from the top surface of the cantilevers because of the rough surface. Then, the cantilevers emit the light perpendicularly, enabling the brightness change of the device. By the on–off switching of driving voltage, the cantilevers work as a pixel of the optical display.

Fabrication Process of the ECDD

As shown in Fig. 7.382, multicantilevers are fabricated from a silicon substrate with a 2-mm-thick thermal oxide layer. The multicantilevers on the substrate are patterned by photolithography and RIE etching, using a Cr layer as a mask for SiO_2 etching. The silicon substrate under the cantilevers is etched in a solution of 15 wt% TMAH (tetramethyl ammonium hydroxide) solution at 90 °C. Next, the ITO layer as a transparent electrode is sputtered on the cantilevers. A micrograph of the fabricated cantilevers is also shown in Fig. 7.382, nesting eight cantilevers from both sides of the substrate to increase the effective pixel area. The length and width of the cantilevers are 490 and 20 μm, respectively. Due to residual stress of the ITO layer, the

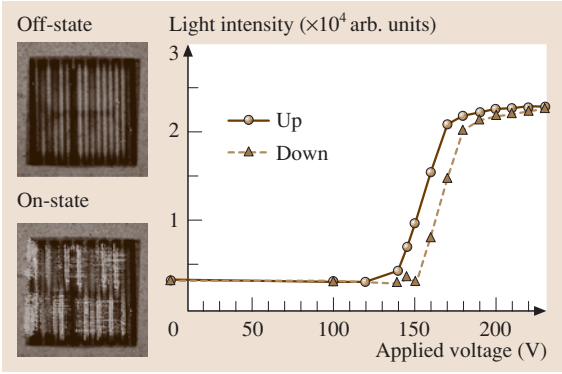


Fig. 7.383 On–off characteristics of the ECDD

fabricated cantilevers are curved toward the silicon substrate with a displacement of 10 μm. Because of the residual stress, the sticking problem between the cantilevers and the substrate is cleared up even after voltage application.

Evaluation of the ECDD

The characteristics of the ECDD were evaluated by the same system as the IDD. A result of the driving evaluation is shown in Fig. 7.383. Although unnecessary scattering is appeared in the off-state, a clear brightness is observed in the on-state of the pixel. A maximum contrast of 0.9, defined by (7.215), was obtained at a driving voltage of 170 V. Hysteresis of the contrast was observed in slow driving test; however, it does not cause the switching speed to deteriorate.

For frequency response, a square voltage of 170 V and a frequency of 1 kHz can drive the multicantilevers. From this performance, the device has a sufficient re-

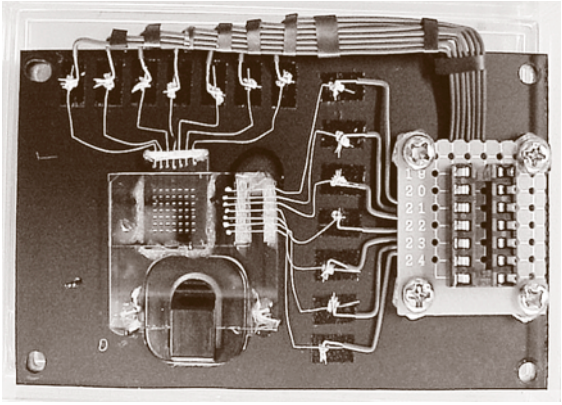


Fig. 7.384 8×8 dot matrix type ECDD

sponse for display application. However, for higher response speed, reduction of the sheet resistance of ITO layers, as well as the size and cantilevers, must be optimized.

A dot-matrix-type device, shown in Fig. 7.384, is fabricated to confirm the function as a display. An 8×8 pixel matrix is constructed in 6×6 mm square, with a pixel size of 500 μm. Using an external voltage driver with a laser incident beam to the glass substrate, the device shows the ability of an individual pixel drive as well as row and column driving.

This type of device can display images on transparent media and can be seen from both sides, when every part is made of transparent material.

References

7.1 DIN: *DIN 8580: Fertigungsverfahren – Bergriffe, Einteilung* (Beuth, Berlin 2003), in German

7.2 K. Herfurth: *Einführung in die Fertigungstechnik* (VEB Verlag Technik, Berlin 1975), in German

7.3 K. Herfurth, N. Ketscher, M. Köhler: *Gießereitechnik kompakt, Werkstoffe, Verfahren, Anwendungen* (Gießerei, Düsseldorf 2003), in German

7.4 W. Hilgenfeld, K. Herfurth: *Tabellenbuch Gusswerkstoffe* (VEB Deutscher Verlag für Grundstoffindustrie, Leipzig 1983), in German

7.5 Guss-Produkte: *Jahreshandbuch für Gußanwender* (Hoppenstedt, Darmstadt 1989), in German

7.6 ZGV: *Feinguß für alle Industriebereiche* (ZGV, Düsseldorf 1984)

7.7 ZGV: *Leitfaden für Gusskonstruktionen* (ZGV, Düsseldorf 1966), in German

7.8 G. Pahl, W. Beitz: *Konstruktionslehre – Handbuch für Studium und Praxis*, 7th edn. (Springer, Heidelberg 2007), in German

7.9 VDI: *VDI-Richtlinie 3237: Fertigungsgerechte Gestaltung von Gusskonstruktionen* (VDI, Düsseldorf 1976), in German

7.10 W. Patterson, R. Döpp: Betriebsnomogramm für Grauguss, Gießerei **47**, 175–180 (1960), in German

7.11 A. Collaud: Gießerei, Tech.-Wiss. Beih. **14**, 709–799 (1954), in German

7.12 N. Ketscher, K. Herfurth, A. Huppertz: Analyse des Energieverbrauchs in Gießereien und Realisierung von Material- und Energieeinsparungen durch Gussteile, Gießerei **88**(1), 21–27 (2001), in German

7.13 EN: *EN 1369: Founding – Magnetic Particle Inspection* (Beuth, Berlin 1997)

7.14 EN: *EN 1370: Founding – Surface Roughness Inspection by Visualtactile Comparators* (Beuth, Berlin 1997)

7.15 EN: *EN 1371-1: Founding – Liquid Penetrant Inspection Part 1: Sand, Gravity Die and Low Pressure Die Castings* (Beuth, Berlin 1997)

7.16 EN: *EN 1371-2: Founding – Liquid Penetrant Inspection Part 2: Investment Castings* (Beuth, Berlin 1998)

7.17 EN: *EN 12454: Founding – Visual Examination of Surface Discontinuities – Steel Sand Castings* (Beuth, Berlin 1998)

7.18 EN: *EN 12680-1: Founding – Ultrasonic Inspection – Part 1: Steel Castings for General Purposes* (Beuth, Berlin 2003)

7.19 EN: *EN 12680-2: Founding – Ultrasonic Inspection – Part 2: Steel Castings for Turbine Components* (Beuth, Berlin 2003)

7.20 EN: *EN 12680-3: Founding – Ultrasonic Inspection – Part 3: Spheroidal Graphite Cast Iron Castings* (Beuth, Berlin 2003)

7.21 EN: *EN 12681: Founding – Radiographic Inspection* (Beuth, Berlin 1998)

7.22 K. Lange (Ed.): *Handbook of Metal Forming* (McGraw-Hill, New York 1985)

7.23 K. Lange (Ed.): *Umformtechnik, Band 1: Grundlagen*, 2nd edn. (Springer, Berlin 1984), in German

7.24 W.D. Callister Jr.: *Materials Science and Engineering* (Wiley, New York 1990)

7.25 ICFG: *ICFG Document No. 11/01: Steels for Cold Forging* (Meisenbach, Bamberg 2001)

7.26 R. Hill: *The mathematical Theory of Plasticity* (Oxford Univ. Press, Oxford 1950)

7.27 L.E. Malvern: *Introduction to the Mechanics of a Continuous Media* (Prentice-Hall, Englewood Cliffs 1969)

7.28 E. Doege, H. Meyer-Nolkemper, I. Saeed: *Fliesskurvenatlas metallischer Werkstoffe* (Hanser, München 1986), in German

7.29 K. Pöhlandt: *Materials Testing for the Metal forming Industry* (Springer, Berlin 1989)

7.30 W.F. Hosford, R.M. Caddell: *Metal Forming. Mechanics and Metallurgy*, 2nd edn. (Prentice-Hall, Englewood Cliffs 1993)

7.31 H. Ismar, O. Mahrenholtz: *Technische Plastomechanik* (Vieweg, Braunschweig 1979), in German

7.32 W. Johnson, P.B. Mellor: *Engineering Plasticity* (Van Nostrand Reinhold, Berkshire 1973)

7.33 R. Kopp, H. Wiegels: *Einführung in die Umformtechnik* (Augustinus, Aachen 1998), in German

7.34 B. Avitzur: *Metal Forming Processes and Analysis* (McGraw-Hill, New York 1968)

7.35 T. Wanheim: Physikalische Prozeßanalyse und -simulation mit nichmetallischen Modellwerkstoffen. In: *Umformtechnik, Band 4*, 2nd edn., ed. by K. Lange (Springer, Berlin 1993), in German

7.36 D. Banabic, H.-J. Bunge, K. Pöhlandt, A.E. Tekkaya: *Formability of Metallic Materials* (Springer, Berlin 2000)
7.37 K. Lange (Ed.): *Umformtechnik, Band 2*, 2nd edn. (Springer, Berlin 1988), in German
7.38 H. Tschätsch: *Praxiswissen Umformtechnik* (Vieweg, Braunschweig 1997), in German
7.39 VDI: *VDI 3171 Kaltmassivumformen, Stauchen, Formpressen* (VDI, Düsseldorf 1958), in German
7.40 K. Lange, H. Meyer-Nolkemper: *Gesenkschmieden* (Springer, Berlin 1977), in German
7.41 S. Kalpakjian, S.R. Schmid: *Manufacturing Engineering and Technology* (Prentice-Hall, New Jersey 2001)
7.42 T. Altan, G. Ngaile, G. Shen: *Cold and Hot Forging. Fundamentals and Applications* (ASM Int., Materials Park 2004)
7.43 T. Altan, S. Oh, H. Gegel: *Metal Forming. Fundamentals and Applications* (ASM Int., Materials Park 1983)
7.44 ICFG: ICFG Document No. 4/82: general aspects of tool design and tool materials for cold and warm forging. In: *ICFG 1967–1992: Objectives, History, Published Documents* (Meisenbach, Bamberg 1992)
7.45 B. Avitzur: *Handbook of Metal Forming Processes* (Wiley, New York 1983)
7.46 A.E. Tekkaya: *Ermittlung von Eigenspannungen in der Kaltmassivumformung* (Springer, Berlin 1986), in German
7.47 G. Spur, T. Stöferle: *Handbuch der Fertigungstechnik. Band 2/1 Umformen* (Hanser, München 1983), in German
7.48 Z. Marciniak, J.L. Duncan, S.J. Hu: *Mechanics of Sheet Metal Forming* (Butterworth-Heinemann, Oxford 2002)
7.49 S.S. Hecker: Cup test for assessing stretchability, Met. Eng. Q. **14**(4), 30–36 (1974)
7.50 K. Lange (Ed.): *Umformtechnik, Band 3*, 2nd edn. (Springer, Berlin 1990), in German
7.51 CIRP: *Dictionary of Production Engineering. Metal Forming 1* (Springer, Berlin 1997)
7.52 Z. Marciniak, J.L. Duncan: *Mechanics of Sheet Metal Forming* (Arnold, London 1992)
7.53 Schuler GmbH: *Metal Forming Handbook* (Springer, Berlin 1998)
7.54 F. Dohmann, C. Hartl: Hydroforming – a method to manufacture light-weight parts, J. Mater. Process. Technol. **60**, 669–676 (1996)
7.55 DIN: *DIN 8589: Manufacturing Processes Chip Removal* (Beuth, Berlin 2003), in German
7.56 H.K. Tönshoff, B. Denkena: *Spanen* (Springer, Berlin 2004), in German
7.57 H.J. Ernst, M.E. Merchant: Chip formation, friction and finish, Trans. Am. Soc. Mech. Eng. **29**, 299–378 (1941)
7.58 DIN: *DIN 8589 Part 1: Manufacturing Processes Chip Removal – Part 1: Turning; Classification, Subdivision, Terms and Definitions* (Beuth, Berlin 2003), in German
7.59 DIN: *DIN 6581: Terminology of Chip Removing; Reference Systems and Angles on the Cutting Part of the Tool* (Beuth, Berlin 1985), in German
7.60 VDEh: Stahl-Eisen-Prüfblatt 1178-69, Verein Deutscher Eisenhüttenleute, in German
7.61 I.S. Jawahir, C.A. van Lutterfeld: Recent developments in chip control research and applications, Ann. CIRP **42**(2), 659–693 (1993)
7.62 DIN: *DIN 6584: Terminology of Chip Removing; Forces, Energy, Work, Power* (Beuth, Berlin 1982), in German
7.63 O. Kienzle, H. Victor: Die Bestimmung von Kräften und Leistungen an spanenden Werkzeugmaschinen, VDI-Zeitschrift **94**, 299–305 (1952), in German
7.64 ISO: *ISO 3685: Tool-Life Testing with Single-Point Turning Tools* (Beuth, Berlin 1993), in German
7.65 F.W. Taylor: On the art of cutting metals, Trans. Am. Soc. Mech. Eng. **28**, 30–351 (1907)
7.66 G. Spur: Beitrag zur Schnittkraftmessung beim Bohren mit Spiralbohrern unter Berücksichtigung der Radialkräfte. Ph.D. Thesis (TU Braunschweig, Braunschweig 1961), in German
7.67 Y. Altintas, M. Weck: Chatter stability of metal cutting and grinding, Ann. CIRP **53**(2), 619–642 (2004)
7.68 J. Tlusty: *Manufacturing processes and Equipment* (Prentice Hall, Eglewood Cliffs 2000)
7.69 F. Klocke, E. Brinksmeier, K. Weinert: Capability profile of hard cutting and grinding, Ann. CIRP **54**(2), 557–580 (2005)
7.70 H.K. Tönshoff, C. Arendt, R. Ben Amor: Cutting of hardened steel, Ann. CIRP **49**(2), 547–566 (2000)
7.71 H.K. Tönshoff, B. Karpuschewski, C. Borbe: Hard machining – state of research, Proc. Int. CIRP/VDI Conf. High Perform. Tools (Düsseldorf 1998) pp. 253–277
7.72 W. König, F. Klocke: *Fertigungsverfahren: Drehen, Fräsen, Bohren* (Springer, Berlin 2007), in German
7.73 M. Klinger: Räumen einsatzgehärteter Werkstücke. Ph.D. Thesis (RWTH Aachen, Aachen 1993), in German
7.74 G.S. Fox-Rabinovich, G.C. Weatherly, A.I. Dodonov, A.I. Kovalev, L.S. Shuster, S.C. Veldhuis, G.K. Dosbaeva, D.L. Wainstein, M.S. Migranov: Nanocrystalline filtered arc deposited (FAD) TiAlN PVD coatings for high-speed machining applications, Surf. Coat. Technol. **177/178**, 800–811 (2004)
7.75 C. Mendibide, P. Steyer, C. Esnouf, P. Goudeau, D. Thiaudière, M. Gailhanou, J. Fontaine: X-ray diffraction analysis of the residual stress state in PVD TiN/CrN multilayer coatings deposited on tool steel, Surf. Coat. Technol. **200**, 165–169 (2005)
7.76 E. Brinksmeier, L. Autschbach: Development of Ultraprecise Milling Techniques for the Manufacture of Optical quality Molds with Continuous and Microstructured Surfaces, Proc. 4th euspen Int. Conf. (Aachen 2003) pp. 193–197

7.77 P. Andrae: Hochleistungszerspanung von Aluminiumknetlegierungen. Ph.D. Thesis (Univ. Hannover, Hannover 2002), in German

7.78 E. Brinksmeier, O. Riemer, R. Stern: Machining of precision parts and microstructures, Proc. 10th Int. Conf. Precis. Eng. (ICPE) (Yokohama 2001) pp. 3–11

7.79 J. Fleischer, B. Denkena, B. Winfough, M. Mori: Workpiece and tool handling in metal cutting machines, Ann. CIRP **55**(2), 817–840 (2006), in German

7.80 M. Weck: *Werkzeugmaschinen – Maschinenarten und Anwendungsbereiche* (Springer, Berlin 1998), in German

7.81 N. Ikawa, R.R. Donaldson, R. Komanduric, W. König, P.A. McKeown, T. Moriwaki, I.F. Stowers: Ultraprecision metal cutting – the past, the present and the future, Ann. CIRP **40**(2), 587–594 (1991)

7.82 T. Masuzawa: State of the art of micromachining, Ann. CIRP **49**(2), 473–487 (2000)

7.83 D. Dornfeld, S. Min, Y. Takeuchi: Recent advances in mechanical micromachining, Ann. CIRP **55**(2), 745–768 (2006)

7.84 E. Uhlmann, K. Schauer: Dynamic load and strain analysis for the optimization of micro end mills, Ann. CIRP **54**(1), 75–78 (2005)

7.85 E. Shamoto, T. Moriwaki: Ultraprecision diamond cutting of hardened steel by applying elliptical vibration cutting, Ann. CIRP **48**(1), 441–444 (1999)

7.86 G. Warnecke, S. Siems: Machining of different steel types at high cutting speeds, Ann. Ger. Acad. Soc. Prod. Eng. **8**(1), 1–4 (2001)

7.87 C. Salomon: German Patent, Patent 523594 (1931), in German

7.88 H. Schulz: *High Speed Machining* (Hanser, München 1996)

7.89 H. Schulz, T. Moriwaki: High speed machining, Ann. CIRP **41**(2), 636–645 (1992)

7.90 R. Ben Amor: Thermomechanische Wirkmechanismen und Spanbildung bei der Hochgeschwindigkeitszerspanung. Ph.D. Thesis (Univ. Hannover, Hannover 2003), in German

7.91 A. Kaldos, A. Boyle, I. Dagiloke: Computer aided cutting process parameter selection for high-speed milling, J. Mater. Process. Technol. **61**, 219–224 (1996)

7.92 D. Brandt: Randzonenbeeinflussung beim Hartdrehen. Ph.D. Thesis (Univ. Hannover, Hannover 1995), in German

7.93 S. Malkin: *Grinding Technology: Theory and Applications of Machining with Abrasives* (Ellis Horwood Ltd., Chichester 1989)

7.94 H.K. Tönshoff, I. Inasaki, B. Karpuschewski, T. Mandrysch: Grinding process achievements and their consequences on machine tools – challenges and opportunities, Ann. CIRP **47**(2), 651–668 (1998)

7.95 E. Brinksmeier, C. Heinzel, M. Wittmann: Friction, cooling and lubrication in grinding, Ann. CIRP **48**(2), 581–598 (1999)

7.96 R. Snoeys: The mean undeformed chip thickness as a basic parameter in grinding, Ann. CIRP **20**, 183–186 (1971)

7.97 J. Webster, M. Tricard: Innovations in abrasive products for precision grinding, Ann. CIRP **53**(2), 597–642 (2004)

7.98 E. Saljé: *Begriffe der Schleif- und Konditioniertechnik* (Vulkan, Essen 1991), in German

7.99 H. Ohmori, T. Nakagawa: Surface grinding of silicon wafers electrolytic in-process dressing, Ann. CIRP **39**(1), 329–332 (1990)

7.100 H.K. Tönshoff, T. Friemuth: Electro contact discharge dressing of diamond wheels for tool grinding, Proc. ICPE Int. Conf. Precis. Eng. (Taipei, 1997) pp. 565–570

7.101 E. Westkämper, H.K. Tönshoff: CBN or CD grinding of profiles, Ann. CIRP **42**(1), 371–374 (1993)

7.102 B. Karpuschewski: Sensors for physical properties. In: *Sensors in Manufacturing*, ed. by H.K. Tönshoff, I. Inasaki (Wiley, Weinheim 2001) pp. 123–142

7.103 I. Inasaki, B. Karpuschewski, H.S. Lee: Grinding chatter – origin and suppression, Ann. CIRP **50**(2), 515–534 (2001)

7.104 W.B. Rowe, J.A. Pettit: Avoidance of thermal damage in grinding and prediction of the damage threshold, Ann. CIRP **37**(1), 327–330 (1988)

7.105 C. Guo, M. Campomanes, D. McIntosh, C. Becze, T. Green, S. Malkin: Optimization of continuous dress creep-feed form grinding process, Ann. CIRP **52**(1), 259–262 (2003)

7.106 F. Klocke, E. Brinksmeier, C.J. Evans, I. Inasaki, T. Howes, H.K. Tönshoff, J.A. Webster, D. Stuff: High-speed grinding – fundamentals and state of the art in Europe, Japan and the USA, Ann. CIRP **46**(2), 715–724 (1997)

7.107 B. Karpuschewski, I. Inasaki: Monitoring systems for grinding processes. In: *Condition Monitoring and Control for Intelligent Manufacturing*, ed. by L. Wang, R.X. Gao (Springer, London 2006) pp. 83–107

7.108 J.F.G. Oliveira, D.A. Dornfeld: Application of AE contact sensing in reliable grinding monitoring, Ann. CIRP **50**(1), 217–220 (2001)

7.109 H.K. Tönshoff, B. Karpuschewski, A. Türich: Tooth profile grinding of precision forged gears, 4th World Congr. Gearing and Power Transm., Vol. 2 (Paris 1999) pp. 1697–1708

7.110 T. Brockhoff, E. Brinksmeier: Grind-hardening: a comprehensive view, Ann. CIRP **48**(1), 255–259 (1999)

7.111 M. Lewis: Blasting through today's hard material – ultrasonic cutting matures into a viable machining process, Am. Mach. **146**(6), 42 (2002)

7.112 P. Dennis: Hochleistungsbandschleifen. Ph.D. Thesis (Univ. Hannover, Hannover 1989), in German

7.113 W. König, H.K. Tönshoff, J. Fromlowitz, P. Dennis: Belt grinding, Ann. CIRP **35**, 487–494 (1986)

7.114 E. Brinksmeier, V. Gehring: Automated finishing of dies and molds by belt grinding, 2nd Int. Conf. Die Mold Technol. (Singapore 1992) pp. 78–91

7.115 H. Mushardt: Modellbetrachtungen und Grundlagen zum Innenrundhonen. Ph.D. Thesis (TU Braunschweig, Braunschweig 1986), in German

7.116 U.-P. Weigmann: Honen keramischer Werkstoffe. Ph.D. Thesis (TU Berlin, Berlin 1997), in German

7.117 E. Saljé, M. von See: Process-optimization in honing of automotive cylinders, Ann. CIRP **36**(1), 235–238 (1987)

7.118 H.K. Tönshoff, C. Marzenell: Properties of tooth surfaces due to gear honing with electroplated tools, 4th World Congr. Gearing Power Transm. (Paris, 1999) pp. 1711–1724

7.119 M. Hartmann: Stabstirn-Trennschleifen von einkristallinem Silizium. Ph.D. Thesis (Univ. Hannover, Hannover 1997), in German

7.120 H.K. Tönshoff, H.-G. Wobker, M. Klein, C. Menz: Precision grinding and slicing of Si-wafers, 7th Int. Precis. Eng. Semin. (Kobe 1993)

7.121 C.J. Evans, E. Paul, D. Dornfeld, D.A. Lucca, G. Byrne, M. Tricard, F. Klocke, O. Dambon, B.A. Mullany: Material removal mechanisms in lapping and polishing, Ann. CIRP **52**(2), 611–634 (2003)

7.122 G. Spur, D. Simpfendörfer: Neue Erkenntnisse und Entwicklungstendenzen beim Planläppen. In: *Jahrbuch Schleifen, Honen, Läppen und Polieren*, 55th edn., ed. by E. Saljé (Vulkan, Essen 1988) pp. 469–480, in German

7.123 H.-H. Nölke: Spanende Bearbeitung von Siliziumnitrid-Werkstoffen durch Ultraschall-Schwingläppen. Ph.D. Thesis (University Hannover, Hannover 1980), in German

7.124 A.M. Hoogstrate, B. Karpuschewski: Modelling of the abrasive waterjet cutting process in a modular way, 16th Int. Conf. Water Jetting (Aix en Provence 2002) pp. 139–150

7.125 A.M. Hoogstrate, T. Susuzlu, B. Karpuschewski: High performance cutting with abrasive waterjets beyond 400 MPa, Ann. CIRP **55**(1), 339–342 (2006)

7.126 W. König, F. Klocke: *Fertigungsverfahren: Abtragen, Generieren und Lasermaterialbearbeitung* (Springer, Berlin 2006), in German

7.127 J. Meijer, A. Du, A. Gillner, D. Hoffmann, V.S. Kovalenko, T. Masuzawa, A. Ostendorf, R. Poprawe, W. Schulz: Laser machining by short und ultrashort pulses, state of the art and new opportunities in the age of the photons, Ann. CIRP **51**(2), 531–550 (2002)

7.128 P.M. Lonardo, A.A. Bruzzone: Effect of flushing and electrode material on die sinking EDM, Ann. CIRP **48**(1), 123–127 (1999)

7.129 J.P. Kruth, L. Steven, L. Froyen, B. Lauwers: Study of the white layer of a surface machined by die sinking electro discharge machining, Ann. CIRP **44**(1), 169–172 (1995)

7.130 T. Kawakami, M. Kunieda: Study on factors determining limits of minimum machinable size in micro EDM, Ann. CIRP **54**(1), 167–170 (2005)

7.131 T. Masuzawa, M. Kimura: Electrochemical surface finishing of tungsten carbide alloy, Ann. CIRP **40**(1), 199–202 (1991)

7.132 A. de Silva, H. Altena, J. McGeough: Influence of electrolyte concentration on copying accuracy of precision ECM, Ann. CIRP **52**(1), 165–168 (2003)

7.133 T. Moriwaki, E. Shamoto, K. Inoue: Ultraprecision ductile cutting of glass by applying ultrasonic vibration, Ann. CIRP **41**(1), 141–144 (1992)

7.134 K. Egashira, T. Masuzawa: Micro-ultrasonic machining by the application of workpiece vibration, Ann. CIRP **48**(1), 131–134 (1999)

7.135 V.O. Bushma, V.M. Borovik, R.V. Rodiakina: *Physical Bases of Generating Concentrated Energy Streams* (MEI, Moscow 1999), p. 104

7.136 V.C. Golubev, F.V. Lebedev: *Physical Bases of Technological Lasers* (Vicshya Shkola, Moscow 1987), p. 192

7.137 M. Lucas, J.N. Petzing, A. Cardoni, L.J. Smith: Design and characterisation of ultrasonic cutting tools, Ann. CIRP **50**(1), 149–152 (2001)

7.138 K.L. Kuo: Experimental investigation of brittle material milling using rotary ultrasonic machining, Proc. 35th Int. MATADOR Conf. (Springer, Berlin 2007) pp. 195–198

7.139 Z.W. Zhong, G. Lin: Ultrasonic assisted turning of an aluminum-based metal matrix composite reinforced with SiC particles, Int. J. Adv. Manuf. Technol. **27**(11/12), 1077–1081 (2006)

7.140 N. Rykalin, A. Uglov: *Laser Machining and Welding* (Elsevier, Amsterdam 1980)

7.141 W.M. Steen, K. Watkins: *Laser Material Processing* (Springer, Berlin 2003)

7.142 K.H. Grote, J. Feldhusen: *Dubbel Taschenbuch für den Maschinenbau* (Springer, Berlin 2007), in German

7.143 J.J. Ramsden, D.M. Allen, D.J. Stephenson, J.R. Alcock, G.N. Peggs, G. Fuller, G. Goch: The design and manufacture of biomedical surfaces, Ann. CIRP **56**(2), 687–711 (2007)

7.144 G. Smolka, W. Gillner, L. Bosse, R. Lützeler: Micro electron beam welding and laser machining – potentials of beam welding methods in the microsystem technology, Microsyst. Technol. **10**(3), 187–192 (2004)

7.145 G. Spur, T. Stöferle: *Handbuch der Fertigungstechnik – Abtragen, Beschichten* (Fachbuchverlag, Leipzig 1998), in German

7.146 J.R. Duncan: Electrochemical grinding of a stainless steel felt, J. Appl. Electrochem. **6**(3), 275–277 (1976)

7.147 E. Uhlmann, S. Piltz, U. Doll: Electrical discharge grinding (EDG) using microstructured disk electrodes, Ann. Ger. Acad. Soc. Prod. **8**(1), 25–31 (2001)

7.148 T. Masuzawa, M. Fujino, K. Kobayashi: Wire-electro-discharge grinding for micro-machining, Ann. CIRP **34**(1), 431–434 (1985)

7.149 A.K. Zaboklicki: Laserunterstütztes Drehen von dichtgesinterter Siliciumnitrid-Keramik. Ph.D. Thesis (RWTH Aachen, Aachen 1998), in German

7.150 W. Moos, R. Janßen-Timmen, H.-K. Starke: *Macroeconomic and sectoral value added by the production and application of welding technology* (Rheinisch-Westfälisches Institut für Wirtschaftsforschung, Essen 2001), in German

7.151 UN: *World Robotics 2004* (United Nations Publications, Geneva 2004)

7.152 W. Pollmann, D. Radaj: *Simulation of Joining Technologies – Potentials and Limits* (DVS Report 214, Düsseldorf 2001), in German

7.153 R. Killing: Weldability of components made of metallic materials, Praktiker **9**, 348–349 (2000)

7.154 DVS: *Repair Welding on Road Vehicles*, Specialist Books on Welding Technology, Vol. 92 (DVS, Düsseldorf 2001), in German

7.155 D. von Hofe, K. Middeldorf: Innovations in joining technology – processes and products for the future, Paton Weld. J. **9/10**, 149–156 (2000)

7.156 W. Moos, R. Janßen-Timmen: *Macroeconomic and Sectoral Value Added by the Production and Application of Welding Technology* (DVS, Düsseldorf 2005), in German

7.157 U. Dilthey: *Laserstrahlschweißen – Prozesse, Werkstoffe, Fertigung und Prüfung* (DVS, Düsseldorf 2000), in German

7.158 K. Mann, J. Hutfless, A. Ruß: Mit dem Scheibenlaser zu neuen Anwendungen, Stuttgarter Lasertage (2003) pp. 71–75, in German

7.159 A. Ruß, W. Gref, M. Leimser, F. Dausinger, H. Hügel: High speed welding of metal sheets with thin disk Laser, Proc. 2nd Int. WLT-Conf. Lasers Manuf. (2003)

7.160 K. Debschütz, W. Becker, R. Bernhardt, K. Mann: New laser application potential through robot-guided remote laser welding, 3rd Eur. Conf. Exhib. (Bad Nauheim 2002)

7.161 T. Graf: Entwicklungsperspektiven verschiedener Hochleistungslaserkonzepte, Stuttgarter Lasertage (2003) pp. 59–61, in German

7.162 M. Seguchi, S. Fujikawa, K. Furuta, Y. Takenaka, K. Yasui: 1 kW highbeam quality and highly efficient diode-pumped Nd:YAG rod laser, Proc. SPIE **4831**, 101–103 (2003)

7.163 S. Jerems, P. Kaupp, F. Lehleuter, E. Meiners: Innovative CO_2-Laser-Verfahren, Stuttgarter Lasertage (2003) pp. 49–52, in German

7.164 W. Gref, M. Leimser, F. Dausinger, H. Hügel: Vom Doppelfokus zur Fokusmatrix, Stuttgarter Lasertage (2003) pp. 189–192, in German

7.165 D. Lindenau, G. Ambrosy, P. Berger, H. Hügel: Magnetisch beeinflusstes Laserstrahlschweißen, Stuttgarter Lasertage (2001) pp. 40–52, in German

7.166 F. Faisst: Fügetechniken für den Werkstoff Aluminium, EUROFORUM-Konferenz, Herausforderung Aluminium-Industrie (Frankfurt 1999), in German

7.167 W. Gref, A. Ruß, M. Leimser, F. Dausinger, H. Hügel: Double focus technique – influence of the focal distance and intensity distribution an the welding process, Proc. Int. Congr. Laser Adv. Mater. Process. LAMP (Osaka 2002)

7.168 M. Kern, P. Berger, H. Hügel: Magnetisch gestütztes Laserstrahlschweißen, Stuttgarter Lasertage (1999) pp. 12–17, in German

7.169 R. Holtz, M. Jokiel: Neue Strategien beim Mikroschweißen mit gepulsten Lasern, Stuttgarter Lasertage (2003) pp. 203–209, in German

7.170 Fraunhofer-Gesellschaft Institut zur Förderung der angewandten Forschung e.V.: Verfahren zum Bearbeiten von Werkstücken mit Laserstrahlung, insbesondere zum Laserstrahlschweißen, Patent DE 4308971A1 (1993), in German

7.171 M. Kogel-Hollacher, C. Dietz, M. Müller, T. Nicolay: Überwachungs- und Regelungsmethoden für das Laserstrahlschweißen, Stuttgarter Lasertage (1999) pp. 54–55, in German

7.172 M. Müller, J. Müller: Prozessüberwachung beim Laserstrahlschweißen – optische Messmethoden für die industrielle Anwendung, Stuttgarter Lasertage (2003) pp. 135–137, in German

7.173 N. Beier, D. Ditzinger: Laserstrahlschneiden oder Stanzen – Kriterien für den Entscheidungsprozess, Int. Conf. Cutting Technology (Hannover 2002) pp. 25–30, in German

7.174 C. Schnitzel, J. Giesekus: Stab, Scheibe oder Slab – Diodengepumpte Festkörperlaser im Vergleich, Laser-Praxis **2**, 18–21 (2001), in German

7.175 F.O. Olsen: Laser Cutting – Trends in the development, Int. Conf. Cut. Technol. (Hannover 2002) pp. 73–78

7.176 W. O'Neill: Entwicklungen zum Dickblech-Laserschneiden, Internationale Schneidtechnische Tagung ICCT (Hannover 2002) pp. 86–92, in German

7.177 R. Hancock: Laser *turbocharges* oxygen cutting of steel slabs, Weld. J. **8**, 46s–47s (2003)

7.178 L. Abram: Laserschneiden von 30 mm Edelstahl im JobShop, Internationale Schneidtechnische Tagung ICCT (Hannover 2002) pp. 93–96, in German

7.179 T. Schüning: Verbesserung der Schnittfugenbildung beim Laserstrahlschneiden durch Erhöhung der Impulsübertragung aus Schneidstrahlen. Ph.D. Thesis (Shaker, Aachen 2002), pp. 1–118, in German

7.180 G. Luxenburger, A. Delahaye, A. Demmerath: Optimierung der Laserschneideignung von Grobblechen, DVS Ber. **220**, 143–147 (2002), in German

7.181 C. Föhl, D. Breitling, F. Dausinger: Präzisionsbohren von Metallen und Keramiken mit kurz- und ultrakurzgepulsten Festkörperlasern, Stuttgarter Lasertage (2003) pp. 91–95, in German

7.182 L. Mayor: Wenn sich Wasser und Feuer verbünden, Schweizer Maschinenmarkt **104**, 40–42 (2003), in German

7.183 B. Richerzhagen: Das Beste von beiden – Laser und Wasserstrahl in einem Prozess kombiniert: Der wassergeführte Laser, Internationale Schneidtechnische Tagung ICCT 2002 (Hannover 2002) pp. 195–202, in German

7.184 K. Dickmann, F. von Alvensleben, S. Friedel: Fein- und Mikrobohrungen mit Nd:YAG-Q-Switch-Laser hoher Strahlqualität, Laser Optoelektron. **6**, 56–62 (1991), in German

7.185 M. von Allmen, A. Blatter: *Laser-Beam Interaction with Materials* (Springer, Berlin 1995), in German

7.186 E. Meiners: *Phänomenologische Untersuchungen zum Bohren von Metallen* (Institut für Strahlwerkzeuge Stuttgart, Stuttgart 1992), Interner Bericht, in German

7.187 F. Lichtner, F. Dausinger: Steuerbare Optik für das Wendelbohren, Laser Mag. **6**, 24–25 (2002), in German

7.188 H. Rohde: *Qualitätsbestimmende Prozessparameter beim Einzelpulsbohren mit einem Nd:YAG-Slablaser* (Teubner, Stuttgart 1999), in German

7.189 G. Bostanjoglo, I. Sarady, T. Beck, G. Phillipps, H. Weber: Bohren von Superlegierungen mit einem gütegeschalteten Nd:YAG-Laser, Laser Optoelektron. **6**, 47–51 (1995), in German

7.190 S. Settegast, T. Beck, C. Föhl, S. Sommer: Bohren im Turbinenbau, Stuttgarter Lasertage (2003) pp. 99–102, in German

7.191 F. Dausinger: Prozessverständnis als Grundlage der Verfahrensentwicklung, Stuttgarter Lasertage (2003) pp. 65–69, in German

7.192 D. Leidinger, R. Holtz, D. Wagner: Applikationen mit mobilen gepulsten Nd:YAG-Laserquellen, Fachtagung in SLV-Halle (2000), in German

7.193 U. Dürr: Gepulste Nd:YAG-Laser im Fahrzeugbau, Laser-Praxis **6**, 34–35 (2000), in German

7.194 H.K. Tönshoff, F. von Alvensleben: *Abtragen und Bohren mit Festkörperlasern* (VDI, Düsseldorf 1993), in German

7.195 S. Schiller, U. Heisig, S. Panzer: *Elektronenstrahltechnologie* (Wissenschaftliche Verlagsgesellschaft, Stuttgart 1977), in German

7.196 H. Schultz: *Elektronenstrahlschweißen* (DVS, Düsseldorf 2000), in German

7.197 U. Dilthey, W. Behr: Elektronenstrahlschweißen an Atmosphäre, Schweiss. Schneid. **52**, 461–465 (2000), in German

7.198 U. Draugelates, B. Bouaifi, B. Ouaissa: Hochgeschwindigkeits-Elektronenstrahlschweißen von-Aluminiumlegierungen unter Atmosphärendruck, Schweiss. Schneid. **52**, 333–339 (2000), in German

7.199 W. Behr: Elektronenstrahlschweißen an Atmosphäre. Ph.D. Thesis (Shaker, Aachen 2003), in German

7.200 J.W. Elmer, A.T. Teruya: An enchanced Faraday cup for rapid determination of power density distribution in electron beams, Weld. J. **80**, 288s–295s (2001)

7.201 K.S. Akopiants: System of diagnostics of electron beam in installations for electron beam welding, Paton Weld. J. **10**, 27–30 (2002)

7.202 F.-W. Bach, A. Szelagowsky, R. Versemann, M. Zelt: Non vacuum electron beam welding of light sheet metals and steel sheets, IIW Document Nr. IV-823-02 (2002)

7.203 U. Dilthey, M. Ahmadian, J. Weiser: Strahlvermessungssystem zur Qualitätssicherung beim Elektronenstrahlschweißen, Schweiss. Schneid. **44**(4), 191–194 (1992), in German

7.204 U. Dilthey, S. Böhm, M. Dobner, G. Träger: Comparability and replication of the electron beam welding technology using new tools of the DIABEAM measurement device, EBT '97, 5th Int. Conf. Electron Beam Technol. (Varna 1997) pp. 76–83

7.205 U. Dilthey, A. Brandenburg, M. Schleser: Dispensing and application of unfilled adhesives in the micro range, Weld. Cut. **3**(4), 250–254 (2004)

7.206 U. Dilthey: *Schweißtechnische Fertigungsverfahren*, Vol. 1 (VDI, Berlin 1994), in German

7.207 W.M. Steen, M. Eboo: Arc augmented laser welding, Met. Construct. **11**, 332–335 (1979)

7.208 U. Dilthey, F. Lüder, A. Wieschemann: Laserstrahlschweißen in der Fertigung – Einsatz und Entwicklung, Bänder Bleche Rohre **37**(11), 26–39 (1996), in German

7.209 DVS: *M 3216 – Laserstrahl-Lichtbogen-Hybridschweißverfahren*, Vol. 01/2005 (DVS, Düsseldorf 2005), in German

7.210 N. Abe, M. Hayashi: Trends in laser arc combination welding methods, Weld. Int. **16**(2), 94–98 (2002)

7.211 C. Walz, I. Stiebe-Springer, M. El Rayes, T. Seefeld, G. Sepold: Hybrid welding of steel for offshore applications, Proc. 11th Int. Offshore Polar Eng. Conf. Exhib., ISOPE 2001 (Stavanger 2001) pp. 263–266

7.212 A. Wieschemann: Entwicklung des Hybrid- und Hydraschweißverfahrens am Beispiel des Schiffbaus. Ph.D. Thesis (RWTH Aachen, Aachen 2001), in German

7.213 U. Dilthey, M. Dobner, A. Ghandehari, F. Lüder, G. Träger: Entwicklung, Stand und Perspektiven der Strahltechnik, 4th Conf. Strahltech. (Halle 1996) pp. 1–13, in German

7.214 D. Petring, S. Kaierle, M. Dahmen, M. Kasimir, F. Cottone, C. Maier: Erweitertes Anwendungsspektrum des Laserstrahlschweißens durch Laser-MIG-Hybridtechnik, Laser Optoelektron. **33**(1), 50–56 (2001), in German

7.215 C. Maier, P. Reinhold, H. Maly, K. Behler, E. Beyer, N. von Heesen: Aluminium-Strangpreßprofile im Schienenfahrzeugbau, geschweißt mit dem Hybridverfahren Nd:YAG-Laser/MIG, DVS Ber. **176**, 198–202 (1996), in German

7.216 A. Keller: CO_2-Laserstrahl-MSG-Hybridschweißen von Baustählen im Blechdickenbereich von 12 bis 15 mm. Ph.D. Thesis (RWTH Aachen, Aachen 2001), in German

7.217 K. Behler, C. Maier, A. Wieschemann: Kombiniertes Laser-Lichtbogenschweißen – Erweiterungspotential für die Lichtbogentechnik, Aachener Schweißtechnik Kolloquium, ASTK'97 (Shaker, Aachen 1997) pp. 151–172, in German

7.218 U. Dilthey, F. Lüder: Untersuchungen zum Laserstrahlschweißen hochkohlenstoffhaltiger Stähle unter Einsatz von Zusatzwerkstoff, Final Rep. DFG-Research Project Di 434/18–4 (1997), in German

7.219 U. Dilthey, M. Biesenbach: Untersuchung der Randbedingungen für die Bildung von *acicular ferrite* in Schweißgütern bei schneller Abkühlung, Final rep AiF-Research Project 11.377 N (2000), in German

7.220 M. Kutsuna, L. Chen: Research on laser-MAG hybrid welding of carbon steel, 7th Int. Weld. Symp. Jap. Weld. Soc. (Kobe 2001) pp. 403–408, in German

7.221 M. Yoneda, M. Katsumura: Laser Hybrid Processing, J. Jpn. Weld. Soc. **58**(6), 427–434 (1989)

7.222 M. Hamasaki: Welding method taking laser welding and MIG welding, Jpn. Patent JP 5966991 (1984), see http://www.nas.gov.ua/pwj/books/lasarc_r.html No. 24

7.223 M. Hamasaki: Welding method combining laser welding and mig welding, US Patent 4507540 (1985)

7.224 Y. Makino, K. Shiihara, S. Asai: Combination welding between CO_2 laser beam and MIG arc, Weld. Int. **16**(2), 99–103 (2002)

7.225 C. Maier: Laserstrahl-Lichbogen-Hybridschweißen von Aluminiumwerkstoffen. Ph.D. Thesis (RWTH Aachen, Aachen 1999), in German

7.226 U. Draugelates: Untersuchungen zur Entwicklung einer plasmalichtbogengestützten Laserstrahltechnik, Project Delineation FA6 (1995), in German

7.227 F. Roland, H. Lembeck: Laserschweißen im Schiffbau – Erfahrungen und Perspektiven auf der Meyer Werft, 7th Int. Aachener Schweißtechnik Kolloquium (iASTK) (Aachen 2001) pp. 463–475, in German

7.228 K. Behler, J. Berkmanns, E. Beyer, Y. Meyer, B. Winderlich: Laserstrahlgeschweißte maßgeschneiderte Bleche aus Aluminium für die industrielle Fertigung, DVS Ber. **170**, 266–272 (1995), in German

7.229 O. Hahn, U. Klemens: *Fügen durch Umformen* (Studienges Stahlanwendung, Düsseldorf 1996), in German

7.230 DIN: *DIN 8593-5: Fertigungsverfahren Fügen – Teil 5: Fügen durch Umformen, Einordnung, Unterteilung, Begriffe* (Beuth, Berlin 2003), in German

7.231 F. Riedel: Möglichkeit der Optimierung von punktförmigen, kraft- und formschlüssigen Feinblechverbindungen am Beispiel Clinchverbindungen und Clinchkonstruktionen. Ph.D. Thesis (Univ. Chemnitz, Chemnitz 2004), in German

7.232 G. Spur: *Handbuch der Fertigungstechnik, Band 5* (Hanser, Munich 1986), in German

7.233 DVS/EFB: *Merkblatt 3420: Clinchen Überblick* (Beuth, Berlin 2002), in German

7.234 M. Todtermuschke: Verfahrensoptimierung zur Herstellung einer punktförmigen, mechanisch gefügten, einseitig ebenen Verbindung ohne Verbindungselement. Ph.D. Thesis (Univ. Chemnitz, Chemnitz 2006), in German

7.235 K.-J. Matthes, F. Riedel (Ed.): *Fügetechnik* (Fachbuchverlag, Leipzig 2003), in German

7.236 J. Grandt: *Blindniettechnik* (VMI, Landsberg 1994), in German

7.237 L. Budde, R. Pilgrim: *Stanznieten und Durchsetzfügen* (VMI, Landsberg 1995), in German

7.238 DVS/EFB: *Merkblatt 3410: Stanznieten* (Beuth, Berlin 2002), in German

7.239 J. Grandt: *Schließringbolzensysteme* (VMI, Landsberg 2001), in German

7.240 DVS/EFB: *Merkblatt 3480: Prüfung von Verbindungseigenschaften – Prüfung der Eigenschaften mechanisch und kombiniert mittels Kleben gefertigter Verbindungen* (Beuth, Berlin 2007), in German

7.241 W. Menz, J. Mohr: *Mikrosystemtechnik für Ingenieure* (VCH, Weinheim 1997), in German

7.242 E. Lugscheider, S. Ferrara: Filler metals for microjoints: New developments for the processes active soldering and transient liquid phase bonding, DVS Rep. **231**, 281–284 (2004)

7.243 S. Böhm: *Fügen in der Feinwerk- und Mikrotechnik* (IFS, Braunschweig 2007), http://www.ifs.-reak tu-braunschweig.de, in German

7.244 K. Lindner: Parallel gap welding, a method of resistance welding to bond fine wires and ribbons, DVS Rep. **124**, 113–117 (1989)

7.245 S. Reul, W. Snakker: Parallel gap and ultrasonic welding at space solor generators – methods, temperature measurment, finite element simulation and low cycle fatigue, DVS Rep. **124**, 118–122 (1989)

7.246 K. Schade: *Mikroelektroniktechnologie* (Verlag Technik, Berlin 1991), in German

7.247 J. Wodara: *Ultraschallfügen und -trennen* (DVS, Düsseldorf 2004), in German

7.248 Small Precision Tools, Lyss, http://www.smallprecisiontools.com

7.249 C.J. Daves, K.I. Johnson, M.H. Scott: Ultrasonic Ball/Wedge Bonding of Aluminium Wires, Electrocomponent Science and Technology **7**, 119–124 (1980)

7.250 G. Schmitz, K. Lindner: *Mikroverbindungstechnik* (DVS, Düsseldorf 2008), in German, http://www.dvs-ev.de/fv/neu/aktuell/Vortrag/Mikroverbindungstechnik_GST_2003.pdf

7.251 K. Lindner: Mikrofügen – Stand der Technik und Trends. In: *Jahrbuch Schweißtechnik 2005* (DVS, Düsseldorf 2004) pp. 129–138, in German

7.252 R.J. Klein Wassink: *Soldering in Electronics*, 2nd edn. (Electrochemical Publications, Ayr 1989)

7.253 J. Zell: Weichlöten in der Elektronik. In: *Jahrbuch Schweißtechnik 1989* (DVS, Düsseldorf 1988) pp. 283–290, in German

7.254 S. Wege, T. Lauer: Das Prozessverhalten der bleifreien Lotlegierungen SnCu für das Wellenlöten und SnAgCu für das Reflowlöten, DVS Ber. **227**, 55–63 (2003), in German

7.255 W.-J. Fischer: *Mikrosystemtechnik* (Vogel, Würzburg 2000), in German

7.256 F. Frisch: Nanotechnologie beflügelt die Klebtechnik, Adhäsion **47**(4), 16–19 (2003), in German

7.257 U. Dilthey, A. Brandenburg, M. Möller: Study of factors influencing the microdosing of unfilled adhesives, J. Micromech. Microeng. **11**, 474–480 (2001)

7.258 U. Dilthey, A. Goumeniouk, S. Böhm, T. Welters: Electron beam diagnostics: a new release of the DIABEAM system, Vacuum **62**, 77–85 (2001)

7.259 DIN: *DIN 8528: Schweißbarkeit metallische Werkstoffe – Begriffe* (Beuth, Berlin 1973), in German

7.260 D. Rosenthal: Mathematical theory of heat distribution during welding and cutting, Weld. J. **20**(5), 220s–234s (1941)

7.261 D. Rosenthal: The theory of moving sources of heat and its application to metal treatments, Trans. Am. Soc. Mech. Eng. **68**, 849–866 (1946)

7.262 N. Rykalin: *Berechnung der Wärmevorgänge beim Schweißen* (VEB Technik, Berlin 1957), in German

7.263 J. Goldak, A. Chakravarti, M. Bibby: A new finite element model for welding heat sources, Met. Trans. B **15**, 299–305 (1984)

7.264 A. Norman, R. Ducharme, A. Mackwood, P. Kapadia, P. Prangnell: Application of thermal modelling to laser beam welding of aluminum alloys, Sci. Technol. Weld. Join. **3**, 260–266 (1998)

7.265 K. Kondoh, T. Ohji: In process heat input in arc welding, Sci. Technol. Weld. Join. **3**, 295–303 (1998)

7.266 M. Hermans, G. den Ouden: Modelling of heat transfer in short circuiting gas metal arc welding, Sci. Technol. Weld. Join. **3**, 135–138 (1998)

7.267 R. Suzuki, O. Trevisan, R. Trevisan: Analytical solutions for heat flow in multiple pass welding, Sci. Technol. Weld. Join. **5**, 63–70 (2000)

7.268 N. Nguyen, A. Ohta, K. Matsuoka, N. Suzuki, Y. Maeda: Analytical solutions for transient temperature of semi-infinite body subjected to 3-D moving heat sources, Weld. J. **78**(8), 265s–274s (1999)

7.269 D. Radaj: *Wärmewirkungen des Schweißens* (Springer, Berlin 1988), in German

7.270 V. Kamala, J. Goldak: Error due to two dimensional approximation in heat transfer analysis of welds, Weld. J. **72**(9), 440s–446s (1993)

7.271 T. Eagar, N. Tsai: Temperature fields produced by traveling distributed heat sources, Weld. J. **62**(12), 346s–355s (1983)

7.272 T. Kasuya, N. Yurioka: Prediction of welding thermal history by a comprehensive solution, Weld. J. **72**(3), 107s–115s (1993)

7.273 S. Jeong, H. Cho: An analytical solution to predict the transient temperature distribution in fillet arc welds, Weld. J. **76**, 223s–232s (1997)

7.274 V.A. Sudnik, A.V. Ivanov, W. Diltey: Mathematical model of a heat source in gas-shielded consumable electrode arc welding, Weld. Int. **15**, 146–152 (2001)

7.275 T. Ninh, N.T. Nguyen, Y.W. Mai, A. Ohta: A new hybrid double-ellipsoidal heat source model for weld pool simulation, Australas. Weld. J. **46**, 39–46 (2001)

7.276 E. Bonifaz: Finite element analysis of heat flow in single-pass arc welds, Weld. J. **79**(5), 121s–125s (2000)

7.277 S. Murugan, T. Gill, P. Kumar, B. Raj, M. Bose: Numerical modelling of temperature distribution during multipass welding of plates, Sci. Technol. Weld. Join. **5**, 208–214 (2000)

7.278 M. Gu, J. Goldak, E. Hughes: Steady state thermal analysis of welds with filler metal addition, Can. Metall. Q. **32**, 49–55 (1993)

7.279 S. Murugan, P. Kumar, T. Gill, B. Raj, M. Bose: Numerical modelling and experimental determination of temperature distribution during manual metal arc welding, Sci. Technol. Weld. Join. **4**, 357–364 (1999)

7.280 G. Little, A. Kamtekar: The effect of thermal properties and weld efficiency on transient temperatures during welding, Comput. Struct. **68**, 157–165 (1998)

7.281 Z. Cai, S. Wu, A. Lu, H. Zhao, Q. Shi: Line Gauss heat source model – an efficient approach for numerical welding simulation, Sci. Technol. Weld. Join. **6**, 84–88 (2001)

7.282 F.M. Zhou, Z.S. Yu, Y.H. Feng, Y.C. Huang, Y.Y. Qian: Numerical analysis of heat transfer process for double sided tungsten inert gas – metal inert gas weld pool, Sci. Technol. Weld. Join. **8**, 76–78 (2003)

7.283 M. Neuhaus: Zum Einfluss der Schrumpfbehinderung auf das thermomechanische Verhalten geschweißter Bauteile. Ph.D. Thesis (Shaker Verlag, Aachen 2005), in German

7.284 D. Dye, S.M. Roberts, A.M. Korsunsky, K.E. James, B. Benn, R.C. Read: Application of low stress low distortion welding to the gas tungsten arc welding of wrought nickel-base alloy C 263. In: *Mathematical Modelling of Weld Phenomena*, Vol. 6 (2002) pp. 751–765

7.285 G. Murakawa, Y. Iuo: Inherent strain as an interface between computational weld mechanics and its industrial application. In: *Mathematical Modelling of Weld Phenomena*, Vol. 4 (1998) pp. 597–619

7.286 R. Gordon: Residual stress and distortion in welded structures – an overview of current US research initiatives, IIW-Doc. XV-878-95 (1995)

7.287 K. Masubuchi: *Analysis of Welded Structures* (Pergamon, New York 1980)

7.288 V.I. Pavlovski, K. Masubuchi: Research in the USSR on residual stresses and distortion in welded structures, Weld. Res. Council. Bull. **388**, 1–62 (1994)

7.289 H. Hänsch: Schweißeigenspannungen und -verformungen, Berechnungsansätze. In: *Jahrbuch*

Schweißtechnik 1995 (DVS, Düsseldorf 1995) pp. 232–244, in German
7.290 S.A. Kuzminov, V.S. Michailov: Determination of transverse shrinkage in multi-pass welding of aluminum alloy sheets and other alloy sheets, Technol. Sudostroj. **5**, 13–22 (1962)
7.291 F. Wörtmann, W. Mohr: Wärmespannungen bei Schweißungen und ihr Einfluss auf die Sicherheit ausgeführter Konstruktionen, Schweizer Bauztg. **100**(19), 143–246 (1932), in German
7.292 R. Malisius: *Schrumpfungen, Spannungen und Risse beim Schweißen* (DVS, Düsseldorf 1957), in German
7.293 K. Satoh, Y. Ueda, H. Kihara: Recent trend of researches on restraint stresses and strains for weld cracking, IIW-Doc. IX-788-72/X-659-72 (Welding Research Institute, Osaka University, Osaka 1972)
7.294 K. Satoh, Y. Ueda, H. Kihara: Recent trends of research into restraint stresses and strains in relation to weld cracking, Weld. World **11**, 133–156 (1973)
7.295 K. Satoh, Y. Ueda, T. Terasaki: Japanese studies on structural restraint severity in relation to weld cracking (preliminary report), Weld. World **15**, 155–189 (1977)
7.296 M. Watanabe, K. Satoh: Effect of welding conditions on the shrinkage and distortion in welded structures, Weld. J. **40**(8), 377s–384s (1961)
7.297 T. Kannengießer: Untersuchungen zur Entstehung Schweißbedingter Spannungen und Verformungen bei Variablen Einspannbedingungen im Bauteilschweißversuch. Ph.D. Thesis (Shaker, Aachen 2000), in German
7.298 W. Gilde: Beitrag zur Berechnung der Querschrumpfung, Schweißtechnik **1**, 10–14 (1957), in German
7.299 L. Capel: Aluminum welding practice, Br. Weld. J. **5**, 245–248 (1961)
7.300 W. Spraragen, W.G. Ettinger: Shrinkage distortion in welding, Weld. J. **29**, 323s–335s (1950)
7.301 E. Richter, G. Georgi: Nahtquerschnitt und Schrumpfung, ZIS-Mitt. **2**, 148–160 (1970), in German
7.302 K. Satoh, S. Matsui: Reaction stress and weld cracking under hindered contraction, IIW-Doc. IX-574-68 (Commission IX, 1968) pp. 353–375
7.303 T. Böllinghaus, T. Kannengießer, M. Neuhaus: Effects of the structural restraint intensity on the stress strain build up in butt joints. In: *Mathematical Modelling of Weld Phenomena*, Vol. 7 (TU Graz, Graz 2004) pp. 651–669
7.304 T. Böllinghaus, H. Hoffmeister, A. Schwager: Calculations of restraint intensities at large offshore steel structures by finite element analysis. In: *Mathematical Modelling of Weld Phenomena*, Vol. 3 (Institute of Materials, London 1997) pp. 624–651
7.305 N.R. Mandal, C.V.N. Sundar: Analysis of welding shrinkage, Weld. J. **76**, 233s–238s (1997)
7.306 J. Goldak, V. Breiguine, N. Dai: Computational weld mechanics – a progress report on ten grand challenges. In: *Trends in Welding Research IV*, ed. by H.B. Smartt, J.A. Johnson, S.A. David (ASM Int., 1996) pp. 5–11
7.307 C.E. Jackson: The science of arc welding, Weld. J. **39**, 129s–140s (1960)
7.308 C.R. Heiple, J.R. Roper: Mechanism for minor element effect on GTA fusion zone geometry, Weld. J. **61**, 97s–102s (1982)
7.309 G.M. Oreper, T.W. Eager, J. Szekely: Convection in arc weld pools, Weld. J. **62**(11), 307s–312s (1983)
7.310 S. Kou, Y.H. Wang: Computer simulation of convection in moving arc weld pools, Met. Trans. A **17**, 2271–2277 (1986)
7.311 T. Zacharia, A.H. Eraslan, D.K. Aidun, S.A. David: Three-dimensional transient model for arc welding process, Met. Trans. B **20**, 645–659 (1989)
7.312 E. Pardo, D.C. Weckman: Prediction of weld pool and reinforcement dimensions of GMA welds using a finite-element model, Met. Trans. B **20**, 937–947 (1989)
7.313 D.C. Weckman: Modelling thermofluids phenomena in arc welds, Proc. 5th Int. Conf. Trends Weld. Res. (ASM, 1999) pp. 3–18
7.314 T. Böllinghaus, H. Schobbert: Nd-YAG laser powder hybrid welding of austenitic stainless steels, 6th Int. Trends Weld. Res. (ASM, 2003) pp. 453–458
7.315 J.C. Metcalfe, M.B. Quigley: Keyhole stability in plasma arc welding, Weld. J. **54**, 401s–405s (1975)
7.316 H.G. Fan, R. Kovacevic: Keyhole formation and collapse in plasma arc welding, J. Phys. D **32**, 2902–2909 (1999)
7.317 G.E. Cook, R.J. Barnett, A.M. Strauss, K. Andersen: Penetration control for gas tungsten arc welding, Int. Conf. Proc. Model. Control Join. Process. (AWS, 1993) pp. 19–26
7.318 Y.H. Xiao, G. den Ouden: Weld pool oscillation during GTA welding of mild steel, Weld. J. **72**, 428s–434s (1993)
7.319 Y.H. Xiao, G. den Ouden: Sensing GTA weld pool geometry by arc voltage signal processing, Weld. Met. Fabr. **64**, 17–20 (1996)
7.320 C.D. Sorensen, T.W. Eagar: Measurement of oscillations in partially penetrated weld pools through spectral anaysis, J. Dyn. Sys. Meas. Control **112**, 463–468 (1990)
7.321 C. Connelly, G.J. Fetzer, R.G. Gann, T.E. Aurand: Reliable welding of HSLA steels by square wave pulsing using an advanced sensing (EDAP) technique. In: *Advances in Welding Technology and Science* (ASM Int., Materials Park 1986) pp. 421–423
7.322 M. Zacksenhouse, D.E. Hardt: Weld pool impedance identification for size measurement and control, J. Dyn. Sys. Meas. Control **105**, 179–184 (1983)
7.323 B. Hu, G. den Ouden: Weld penetration sensing and control during GTA welding using weld pool oscillation, Proc. 5th Int. Conf. Trends Weld. Res. (ASM 1998) pp. 1125–1130

7.324 Q.L. Wang, C.L. Yang, Z. Geng: Separately excited resonance phenomena of the weld pool and its application, Weld. J. **72**, 455s–462s (1993)

7.325 P. Shewmon: *Diffusion in Solids* (McGraw-Hill, New York 1963)

7.326 D.A. Porter, K.E. Easterling: *Phase Transformations in Metals and Alloys*, 2nd edn. (Chapman Hall, London 1997)

7.327 T. Böllinghaus, H. Hoffmeister: Finite element calculations of pre- and postheating procedures for sufficient hydrogen removal in butt joints. In: *Mathematical Modelling of Weld Phenomena*, Vol. 3 (Institute of Materials, London 1997) pp. 726–756

7.328 T. Böllinghaus, H. Hoffmeister, A. Dangeleit: A scatterband for hydrogen diffusion coefficients in micro-alloyed and low carbon structural steels, Weld. World **35**, 83–96 (1995)

7.329 T. Böllinghaus: Modelling of hydrogen diffusion and cracking in steel welds. In: *Mathematical Modelling of Weld Phenomena*, Vol. 5 (Institute of Materials, London 2001) pp. 1019–1060

7.330 P. Sofronis, R.M. McMeeking: Numerical analysis of hydrogen transport near a blunting crack tip, J. Mech. Phys. Solids **37**, 40–50 (1989)

7.331 T. Böllinghaus, T. Kannengießer, C. Jochum, I. Stiebe-Springer: Effect of filler material selection on stress-strain build up and stress corrosion cracking resistance of supermartensitic stainless steel pipeline welds, IIW-Doc. No. II-A-141-04 (NACE, Houston 2002), paper 02061

7.332 V.A. Karkhin, W. Dreutz, N.O. Pavlova, W. Schulz: Effect of low-temperature transformations on residual stress distributions in laser welded joints. In: *Mathematical Modelling of Weld Phenomena*, Vol. 5 (Institute of Materials, London 2001) pp. 597–614

7.333 H.K.D.H. Bhadeshia: Material factors. In: *Handbook of Residual Stress and Deformation of Steel*, ed. by G. Totten, M. Howes, T. Inoue (ASM Int., Materials Park 2002)

7.334 T. Kannengießer, W. Florian, T. Böllinghaus, H. Herold: Effect of weld metal strength and welding conditions on reaction forces and stress distribution of restrained components, Weld. World **45**, 18–26 (2001)

7.335 T. Böllinghaus, H. Hoffmeister: Numerical model for hydrogen assisted cracking, Corrosion **56**, 611–622 (2000)

7.336 T. Böllinghaus, E. Viyanit: Numerical modelling of hydrogen assisted cracking in girth welds of supermartensitic stainless steel pipelines – Report I. In: *Mathematical Modelling of Weld Phenomena*, Vol. 6 (Institute of Materials, London 2002) pp. 839–855

7.337 A.S. Tetelman, A. McEvily: *Fracture of Structural Materials* (Wiley, New York 1967)

7.338 T. Böllinghaus, E. Viyanit, H. Hoffmeister: Numerical modelling of hydrogen assisted cracking in girth welds of supermartensitic stainless steel pipelines – Report II. In: *Mathematical Modelling of Weld Phenomena*, Vol. 7 (TU Graz, Graz 2005) pp. 847–874

7.339 P. Zimmer, T. Böllinghaus, T. Kannengießer: Effects of hydrogen on weld microstructure properties of the high strength steels S 690 Q and S 1100 QL, IIW-Doc. No. II-A-141-04 (2004)

7.340 W.F. Savage, C.D. Lundin, A.H. Aronson: Weld metal solidification mechanics, Weld. J. **44**, 175s–181s (1965)

7.341 W.F. Savage, A.H. Aronson: Preferred orientation in the weld fusion zone, Weld. J. **45**, 85s–89s (1966)

7.342 W.F. Savage, R.H. Hrubec: Synthesis of weld solidification using crystalline organic materials, Weld. J. **51**, 260s–271s (1972)

7.343 P.E. Brown, C.M. Adams Jr.: Rapidly solidified alloy structures, Trans. Am. Foundrymen's Soc. **69**, 879–891 (1961)

7.344 P.E. Brown, C.M. Adams Jr.: Fusion zone structures and properties in aluminum alloys, Weld. J. **39**, 520s–524s (1960)

7.345 G.J. Davies, J.G. Garland: Solidification structures and properties of fusion welds, Int. Met. Rev. **20**, 83–105 (1975)

7.346 S.A. David, J.M. Vitek: Correlation between solidification parameters and weld microstructures, Int. Mater. Rev. **34**, 213–245 (1989)

7.347 C.A. Gandin, M. Rappaz: A coupled finite element-cellular automation model for the prediction of dendrite grain structures in solidification processes, Acta Met. Mater. **42**, 2233–2246 (1994)

7.348 U. Dilthey, T. Reichel, V. Pavlik: A modified cellular automata model for grain growth simulation. In: *Mathematical Modelling of Weld Phenomena*, Vol. 3 (Institute of Materials, London 1997) pp. 106–113

7.349 V.V. Ploshikhin, H.W. Bergmann: Simulation of grain structures in laser beam welds undergoing the plasma solidification mode. In: *Mathematical Modelling of Weld Phenomena*, Vol. 4 (Institute of Materials, London 1998) pp. 150–165

7.350 H.W. Bergmann, S. Mayer, K. Müller, V.V. Ploshikhin: Texture evolution in laser beam welds undergoing the planar solidification mode. In: *Mathematical Modelling of Weld Phenomena*, Vol. 4 (Institute of Materials, London 1998) pp. 166–183

7.351 V.V. Ploshikhin, H.W. Bergmann: Correlation between the welding parameters and the grain structure for the fast moving high power line heat source in a thin plate. In: *Mathematical Modelling of Weld Phenomena*, Vol. 5 (Institute of Materials, London 2001) pp. 269–282

7.352 K. Ichikawa, A. Nogami, T. Koseki, Y. Fukuda: Modelling of solidification and grain growth in steel welds. In: *Mathematical Modelling of Weld Phenomena*, Vol. 5 (Institute of Materials, London 2001) pp. 189–209

7.353 M.H. Burden, J.D. Hunt: Cellular and dendritic growth II, J. Cryst. Growth **22**, 109–116 (1974)

7.354 J.A. Brooks: Weld solidification and microstructural development, 4th Int. Trends Weld. Res. (ASM 1995) pp. 123–134

7.355 J.D. Hunt: Steady state columnar and equiaxed growth of dendrites and eutectic, Mater. Sci. Eng. **65**, 75–83 (1984)

7.356 Ø. Grong, C.E. Cross: A model for predicting weld metal grain refinement in G-V space, Mater. Res. Soc. Symp. Proc. **578**, 431–438 (2000)

7.357 E.Z. Scheil: Bemerkungen zur Schichtkristallbildung, Z. Metallk. **34**, 70–72 (1942), in German

7.358 B. Radhakrishnan, R.G. Thompson: A phase diagram approach to study liquation cracking in alloy 718, Met. Trans. A **22**, 887–902 (1991)

7.359 N.F. Gittos, M.H. Scott: Heat-affected zone cracking of Al-Mg-Si alloys, Weld. J. **60**, 95s–103s (1981)

7.360 M. Katoh, H.W. Kerr: Investigation of heat-affected zone cracking of GTA welds of Al-Mg-Si alloys using the varestraint test, Weld. J. **66**, 360s–368s (1987)

7.361 M. Wolf, H. Schobbert, T. Böllinghaus: Influence of the weld pool geometry on solidification crack formation. In: *Hot Cracking Phenomena in Welds*, ed. by T. Böllinghaus, H. Herol (Springer, Berlin 2005) pp. 245–268

7.362 J.J. Pepe, W.F. Savage: Effects of constitutional liquation in 18-Ni maraging steel weldments, Weld. J. **46**, 411s–422s (1967)

7.363 B. Radhakrishnan, R.G. Thompson: A model for the formation and solidification of grain boundary liquid in the heat-affected zone (HAZ) of welds, Met. Trans. A **23**, 1783–1799 (1992)

7.364 C. Huang, S. Kou: Partially melted zone phenomena in aluminum welds – binary Al-Cu alloys, Conf. Proc. 6th Int. Trends Weld. Res. (ASM, 2003) pp. 633–637

7.365 C. Huang, S. Kou: Liquation cracking in full-penetration Al-Cu welds, Weld. J. **83**, 50s–58s (2004)

7.366 C.E. Cross: On the origin of weld solidification cracking. In: *Hot Cracking Phenomena in Welds*, ed. by T. Böllinghaus, H. Herol (Springer, Berlin 2005) pp. 3–18

7.367 W.G. Savage, C.D. Lundin: The varestraint test, Weld. J. **44**, 433s–442s (1965)

7.368 T.W. Nelson, J.C. Lippold, W. Lin, W.A. Baeslack III: Evaluation of the circular patch test for assessing weld solidification cracking, Part I – Development of a test method, Weld. J. **76**, 110s–119s (1997)

7.369 G.M. Goodwin: Development of a new hot-cracking test – The sigmajig, Weld. J. **66**, 33s–38s (1987)

7.370 H. Herold, M. Streitenberger, A. Pchennikov: Modelling of the PVR-test to examine the origin of different hot cracking types. In: *Mathematical Modelling of Weld Phenomena*, Vol. 5 (Institute of Materials, London 2001) pp. 783–792

7.371 N.N. Prokhorov: The problem of the strength of metals while solidifying during welding, Svar. Proiz. **6**, 5–11 (1956)

7.372 T. Senda, F. Matsuda, G. Takano: Studies on solidification crack susceptibility for weld metals with trans-varestraint test, J. Jpn. Weld. Soc. **42**, 48–56 (1973)

7.373 J.C. Lippold: Recent developments in weldability testing. In: *Hot Cracking Phenomena in Welds*, ed. by T. Böllinghaus, H. Herol (Springer, Berlin 2005) pp. 271–290

7.374 U. Feurer: Influence of alloy composition and solidification conditions on dendritic arm spacing, feeding, and hot tear properties of aluminum alloys, Proc. Int. Symp. Eng. Alloy. (Delft, 1997) pp. 131–145

7.375 J. Campbell: *Castings* (Butterworth-Heinemann, Oxford 1991), pp. 219–229

7.376 M. Rappaz, J.M. Drezet, M. Gremaud: A new hot-tearing criterion, Met. Mater. Trans. A **30**, 449–455 (1999)

7.377 T. Kannengießer, T. McInerney, W. Florian, T. Böllinghaus, C.E. Cross: The influence of local weld derformation on hot cracking susceptibility. In: *Mathematical Modelling of Weld Phenomena*, Vol. 6 (Institute of Materials, London 2002) pp. 803–818

7.378 T. Zacharia: Dynamic stresses in weld metal hot cracking, Weld. J. **73**, 164s–172s (1994)

7.379 Z. Feng, T. Zacharia, S.A. David: On the thermo-mechanical conditions for weld metal solidification cracking. In: *Mathematical Modelling of Weld Phenomena*, Vol. 3 (Institute of Materials, London 1997) pp. 114–148

7.380 J.J. Dike, J.A. Brooks, M. Li: Comparison of failure criteria in weld solidification cracking simulations. In: *Mathematical Modelling of Weld Phenomena*, Vol. 4 (Institute of Materials, London 1998) pp. 199–222

7.381 J. Campbell: Pore nucleation in solidifying metals. In: *The Solidification of Metals* (ISI, London 1968) pp. 18–26

7.382 R.E. Trevisan, D.D. Schwemmer, D.L. Olson: The fundamentals of weld pore formation. In: *Welding-Theory and Practice* (North-Holland, Amsterdam 1990) pp. 79–115

7.383 T.A. Palmer, T. DebRoy: Physical modeling of nitrogen partition between the weld metal and its plasma environment, Weld. J. **75**, 197s–207s (1996)

7.384 K. Kubo, R.D. Pehlke: Mathematical modeling of porosity formation in solidification, Met. Trans. B **16**, 359–366 (1985)

7.385 D.R. Poirier, K. Yeum, A.L. Maples: A thermodynamic prediction for microporosity formation in aluminum-rich Al-Cu alloys, Met. Trans. A **18**, 1979–1987 (1987)

7.386 W.F. Brown, J. Bandas, N.T. Olson: Pulsed magnetic welding of breeder reactor fuel pin end closures, Weld. J. **57**(6), 22s–26s (1978)

7.387 A. Weber: Magnetic pulse technology attracts new users, Assembly Mag. **45**(9), 58–63 (2002)

7.388 E.V. Onosovskii, V.A. Chudakov, V.I. Sokolov, V.D. Saprygin: Magnetic pulse welding of thin-walled aluminum-steel adapters, Kim. Neft. Mashinostr. **11**, 25–26 (1984)

7.389 V.P. Epechurin: Properties of bimetal joints produced by magnetic-pulse welding, Svar. Proiz. **5**, 12–14 (1974)
7.390 E.S. Karakozov, Z.A. Chankvetadze, N.M. Beriev: The interaction of metals in magnetic impulse welding, Svar. Proiz. **12**, 4–6 (1977)
7.391 V.A. Chudakov: The effect of the temperature to which the material is heated on the process of formation of intermetallic compounds in magnetic pulse welding, Svar. Proiz. **9**, 16–18 (1980)
7.392 K.K. Khrenov, V.A. Chudakov: The formation of welded joints in the magnetic pulsed welding of cylindrical workpieces, Weld. Prod. (USSR) **25**(9), 19–20 (1978)
7.393 M. Marya, S. Marya: Interfacial microstructures and temperatures in aluminum–copper electromagnetic pulse welds, Sci. Technol. Weld. Join. **9**(6), 541–547 (2004)
7.394 M. Marya, D. Priem, S. Marya: Microstructures at aluminum–copper magnetic pulse weld interfaces, Proc. THERMEC 2003 Int. Conf. Process. Manuf. Adv. Mater. (Madrid 2003)
7.395 M. Marya, S. Marya, D. Priem: On the characteristics of electromagnetic welds between aluminum and other metals and alloys, IIW Doc. IX–2141–04 (2004)
7.396 L.I. Markashova, Y.U.A. Sergeeva, V.V. Statsenko, V.A. Chudakov: Special features of the mechanism of structure formation in magnetic pulsed welding, Paton Weld. J. **3**(3), 187–191 (1991)
7.397 A. Stern, M. Aizenshtein: Bonding zone formation in magnetic pulse welds, Sci. Technol. Weld. Join. **7**(5), 339–342 (2002)
7.398 V. Shribman, Y. Livshitz, O. Gafri: Magnetic pulse welding and joining – a new tool for the automotive, SAE Technical Paper 2001-01-3408 (2001)
7.399 T. Sano, M. Takahashi, Y. Murakoshi, M. Terasaki, K.I. Matuno: Electromagnetic joining of metal tubes to ceramic rods, J. Jpn. Soc. Technol. Plast. **28**(322), 1193–1198 (1987)
7.400 B. Bourgoin: Le formage électromagnétique, CETIM Inf. **80/81**, 18–26 (1983), in French
7.401 Y. Strizhakov: Calculating and selecting the parameters of magnetic pulsed vacuum welding, Phys. Chem. Mater. Technol. **5**(1), 89–91 (1991)
7.402 M. Kojima, K. Tamaki: Electromagnetic welding of tubes, Proc. 5th Int. Symp. Jpn. Weld. Soc. (1990) pp. 201–206
7.403 H. Baker, H. Okamoto: *ASM Handbook Volume 03, Alloy Phase Diagrams* (American Society for Materials, Pennsylvania 1992)
7.404 K. Ferjutz, J.R. Davis: *ASM Handbook Volume 06, Welding, Brazing and Soldering* (American Society for Materials, Pennsylvania 1993)
7.405 G. Krauss: *Steel, Heat Treatment and Processing Principles* (ASM Int., Materials Park 1990)
7.406 H. Schultz: Informations- und Komunikationstechnik beeinflußt das Rapid Product Development, Ind. Manag. **14** (1998), in German
7.407 R.F. Scholl: VDI-Zeitschrift **141**(9/10) (1999), in German
7.408 E. Chlebus: *Computer Technix CAx in Production Engineering* (WNT, Warsaw 2000), in Polish
7.409 M. Eigner: Requirements with regard to PDM system architecture and functionality – a vendors report, Proc. Product Data Management based on International Standards (Volkswagen AG, Braunschweig 1999)
7.410 C.-O. Bauer: Produkthaftung-Ansprüche an die Konstruktion haben einen Anteil von 70%, Maschinenmarkt **68** (1984), in German
7.411 E. Westkämper: Manufuture – a vision for 2020, Workshop (Hannover 2004)
7.412 A. Gebhardt: *Rapid Prototyping – Werkzeuge für die schnelle Produktentwicklung* (Hanser, München 1996), in German
7.413 J.J. Beaman: *Additive/Subtractive Manufacturing Research and Development in Europe* (World Technology Evaluation Center, Baltimore 2004)
7.414 E. Chlebus: *Innovative Rapid Prototyping – Rapid Tooling Technologies in Product Development* (Centre for Advanced Manufacturing Technologies, Wrocław University of Technology 2003)
7.415 W. Liu, L. Li, K. Kochar: A method for assessing geometrical errors in layered manufacturing. Part 1: error interaction and transfer mechanisms, J. Int. Adv. Manuf. Technol. **14**, 637–643 (1998)
7.416 W. Liu, L. Li, K. Kochar: A method for assessing geometrical errors in layered manufacturing. Part 2: mathematical modelling and numerical evaluation, J. Int. Adv. Manuf. Technol. **14**, 644–650 (1998)
7.417 R. Simmonds: Silikon und Polyurethan im Prototypenbau, Maschinenmarkt **52** (1997), in German
7.418 C.K. Chua, S.M. Chou, T.S. Wong: A study of the state-of-the-art rapid prototyping technologies, Int. J. Adv. Manuf. Technol. **14**, 146–152 (1998)
7.419 D.T. Pham, R.S. Gault: A comparison of rapid prototyping technologies, Int. J. Mach. Tools Manuf. **38**, 1257–1287 (1998)
7.420 K.E. Oczoś: Rapid prototyping – meaning, characteristic of methods and applications, Mechanik **10** (1997), in Polish
7.421 K.E. Oczoś: Progression in additive manufacturing, Mechanik **4** (1999), in Polish
7.422 G. Spur, E. Uhlmann: *Rapid Prototyping – Dubbel Taschenbuch für den Maschinenbau*, 21st edn. (Springer, Berlin 2005), pp. 94–95, in German
7.423 Wohlers: Wohlers Report 2004: Rapid prototyping, tooling and manufacturing state of the industry, Annual Worldwide Progress Report (Wohlers, Fort Collins 2004)
7.424 Wohlers: Wohlers Report 2006: Rapid prototyping, tooling and manufacturing state of the industry, Annual Worldwide Progress Report (Wohlers, Fort Collins 2006)
7.425 K.E. Oczoś: Rapid Prototyping and Rapid Tooling – development of methods and techniques of rapid

manufacture of models, prototypes and small-series products, Mechanik **4** (1998), in Polish
7.426 M. Meindl: Prototypen in Produktentwicklung, Seminarber. IWB **49** (1999), in German
7.427 http://www.3dsystems.com
7.428 Cubital Ltd: *Cubital Facet List Syntax Guide* (Cubital, Raanana)
7.429 Z Corporation: *Z Corporation family of printers Online in Internet* (Z Corporation, Burlington 2004), www.zcorp.com/products/printers.asp
7.430 J. Kowola: Realizing the potential of 3D Printer, Proc. Euro-URapid 2005 (Leipzig 2005)
7.431 http://www.stratasys.com
7.432 C.M. Stotko: E-Manufacturing: Von den Daten zum fertigen Produkt, Proc. Euro-URapid 2005 (Leipzig 2005), in German
7.433 X. Wu: Direct Laser Fabrication, Proc. Seminar Rapid Product Development (CAMT Wrocław University of Technology, Wrocław 2002)
7.434 M. Schellabear, J. Weilhammer: Direktes Metall-Laser-Sintern (DMLS) – Industrielle Anwendung für Rapid Tooling und Manufacturing, Seminarberichte IWB TU Munich Nr. 50, Rapid Manufacturing – Methoden für die reaktionsfähige Produktion (Augsburg 1999), in German
7.435 http://www.eos.info
7.436 http://www.fockeleundschwarze.de/english/fsrd.html
7.437 http://www.mcp-group.de
7.438 http://www.trumpf.com
7.439 http://www.phenix-system.com
7.440 http://www.cicweb.de, http://www.hig-ag.de
7.441 http://www.arcam.com
7.442 http://www.optomec.com
7.443 Reverse Engineering-Technologies for Reverse Engineering, http://www.myb2o.com/myb2ous/ReverseEngineering/Tools/Process/10618.htm#reverse (2001)
7.444 Immersion Corporation, http://www.immersion.com (2001)
7.445 B. Dybała, P. Kolinka: Technologies of reverse engineering in product development, 4th Conf. Prod. Autom. (Wrocław 2003), in Polish
7.446 B. Dybała: Methods of modelling and prototyping of anatomical objects, Proc. Euro-URapid 2005 (Leipzig 2005)
7.447 R. Hermann, M. Hermann: *Tomografia Komputerowa* (Czerwiec 2001), http://www.zdrowie.med.pl/index.phtml
7.448 http://www.aracor.com (2002)
7.449 Cyberware, http://www.cyberware.com/products/index.html (2001)
7.450 LDI, Laser Design Inc., http://www.laserdesign.com (2001)
7.451 Capture 3D Inc., http://www.capture3d.com/html/products.html (2001)
7.452 Inspec Inc., http://www.sms-ct.com (2001)
7.453 Materialise NV: *Materialise Medical* (Materialise NV, Belgium 2002), http://www.materialise.be
7.454 Microscopic Moire Interferometry, http://www.aem.umn.edu/people/faculty/shield/mm.html (2001)
7.455 Photogrammetry, http://www.univie.ac.at/Luftbildarchiv/intro.htm (2001)
7.456 Align Technology, Inc., http:/www.invisalign.com (2001)
7.457 HEK: *Rapid Prototype Tooling* (HEK GmbH, Germany 2001)
7.458 http://www.axson.com (2006)
7.459 http://www.ivf.se
7.460 K.W. Goosen, J.A. Walker, S.C. Arney: Silicon modulator based on mechanically-active antireflection layer with 1 Mb/s capability or fiber-in-the-loop applications, IEEE Photon. Technol. Lett. **6**(9), 1119–1121 (1994)
7.461 J.B. Sampell: Digital micromirror device and its application to projection displays, J. Vac. Soc. Technol. B **12**, 3242–3246 (1994)
7.462 T. Hatsuzawa, T. Oguchi: Application of micromachined SiO_2 film for display devices, 10th Int. Conf. Solid-State Sens. Actuators (1999) pp. 804–807
7.463 T. Oguchi, M. Hayase, T. Hatsuzawa: Driving performance improvement of the interferometric display device (IDD), Optical MEMS 2001 (2001) pp. 107–108
7.464 T. Hatsuzawa, T. Oguchi, M. Hayase: An electrostatic-driven optical switching structure for display device, Optical MEMS 2001 (2001) pp. 149–150
7.465 T. Oguchi, H. Masanori, T. Hatsuzawa: Electrostatically driven micro-optical switching device based on interference of light and evanescent coupling, Proc. SPIE **4902**, 213–220 (2002)
7.466 T. Oguchi, M. Hayase, T. Hatsuzawa: Electrostatically driven display device using evanescent coupling between sheet waveguide and multicantilevers, Optomechatoronic Systems IV, Proc. SPIE **5264**, 134–141 (2003)

8. Measuring and Quality Control

Norge I. Coello Machado, Shuichi Sakamoto, Steffen Wengler, Lutz Wisweh

Considering the incessantly increasing requirements to the quality of products and processes it is necessary to improve a quality-orientated management in all departments of any types of companies and the advantageous application of manufacturing measurement equipment.

In addition to diverse technical requirements are also to consider the requirements of national, international and company-specific norms. The companies must not only fulfill the requirements of the quality, but also the requirements of safety, environment and economy.

As follows some aspects of the manufacturing measurement technology and quality management and their integration into a manufacturing process will be introduced.

Starting with manufacturing geometrical conditions and statements at drawings (nominal state and geometrical limits) the use of measurement equipment and gages to the evaluation of geometric elements will be described. Basic knowledge to measuring standards, uncertainties as well as calibration and measuring instrument inspection will mediates.

Based on physical principles equipment and methods for the registration of measurement values, form- and position deviations and surface characteristics will introduce.

8.1 Quality Management

8.1.1 Quality and Quality Management

Nowadays the quality of products, assemblies and services not only includes the fulfilment of functional requirements by maintaining tolerances. It also includes the fulfilment of numerous requirements such as rendered in parts in Fig. 8.1. In this section some fundamentals of quality management will be described from the multitude of requirements. In Sect. 8.2 some aspects of the requirements of manufacturing measurement technology for the qualification of the geometrical quality of products will be shown.

Among the requirements for organizations involved in quality control, the key concepts of quality management (QM) and total quality management (TQM) include planning, monitoring, and improvement of quality, such as the consideration of representatives and departments relevant to quality, as shown in Fig. 8.2.

8.1.2 Quality Management Methods

To conform to the requirements of modern quality management, nowadays numerous procedures and methods, with many different applications, are available. Fig-

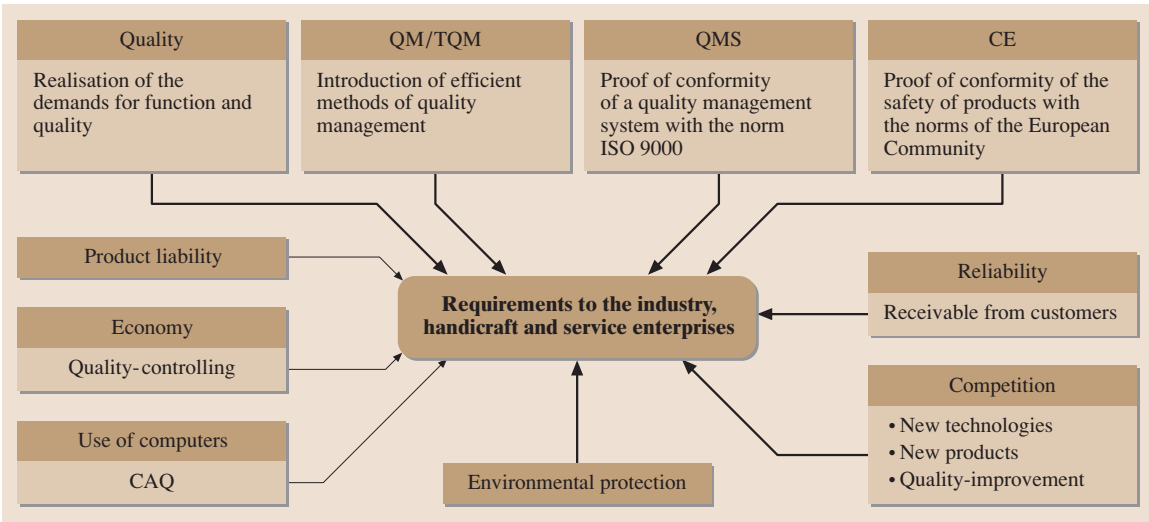


Fig. 8.1 Requirements for industry, craft, and service enterprises

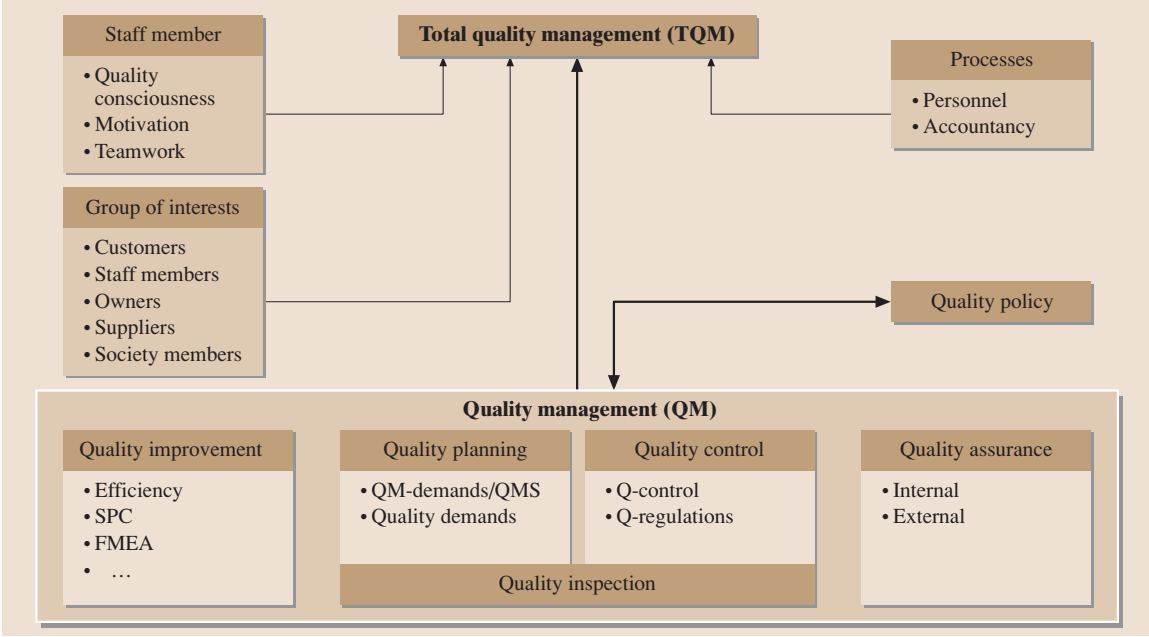


Fig. 8.2 Elements of quality management

ure 8.3 presents a selection of relevant methods used in the area of mechanical engineering.

The different methods can be associated with two essential groups: *problem-resolving techniques* and *preventive techniques*. The problem-resolving techniques can also be divided into the two other categories: those that find causes of existing quality problems and those that help the engineer develop definite aims in a systematically manner (a management method).

The prime conditions for quality analysis are the collection and preparation of quality data with respect to measured data. *Checklists* enable a visualization of the accumulation of certain failures or various kinds of mistakes. *Histograms* demonstrate the probability of

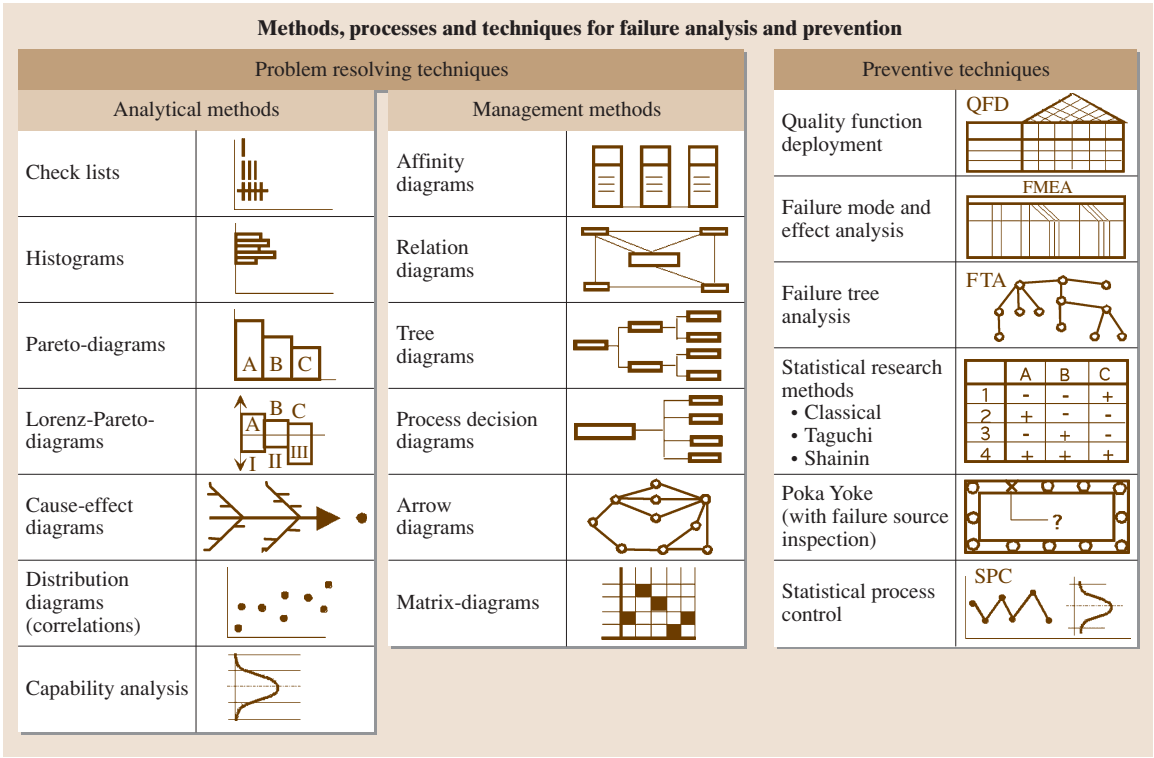


Fig. 8.3 Problem resolving and preventive techniques

the occurrence of a certain event within intervals of the measurement range. *Pareto analyses*, also called *ABC analyses*, enable a weighting of the dominant influencing parameters. In combination with the corresponding costs, they are known as Lorenz–Pareto analyses.

Referring to a definite problem, a comprehensive structuring of clustered influences or causes can be presented as a *cause–effect diagram*, also known as a *fishbone diagram* or *Ishikawa diagram*, which can be used to solve possible problems.

Especially for the experimental aetiology of independent measurements (e.g., quality characteristics) the demonstration of pairs of values in an x–y coordinate system is applicable. These *scatter diagrams* can be described with the help of statistical calculations such as regression and correlation analyses, in which the identified regression function describes the kind of functional connection while the correlation describes the intensity of that connection through the correlation coefficient.

Nowadays an *analysis and evaluation of the quality capability* of procedures and processes takes place in many areas of mechanical engineering by a comprehensive statistical evaluation in terms of capability coefficients, which express the relation of process spread (6σ) to the tolerance and the position of the arithmetic mean of the tolerance borders. In this way it is possible to compare the quality level of different processes.

Management methods support the solution of problems by a targeted systematic procedure. *Affinity diagrams* allow a structured systematization of ideas in order to point out the correlation between these ideas.

Relation diagrams enable the development of cause–effect relations by visualizing networked structures. *Tree diagrams* subdivide specifically aims until directly realizable activities are practicable.

Firstly, *process–decision diagrams* start by arranging possible problems by:

- Urgency
- Probability of occurrence
- Difficulty of prevention

with the aim of detecting potential problems already in the planning phase and to elaborate corresponding countermeasures.

Arrow diagrams or *net plans* are important resources for project planning for the investigation of *critical paths*, which determine the total permanence of a project. In this method the determination of a process sequence is made using series and parallel paths to develop a detailed explanation of the working steps required to achieve the project aim, followed by the assignment of the corresponding process durations.

If a lot of information quantities for certain circumstance are available, *matrix diagrams* are situable for detection of latent structures. By using data evaluation in pairs with the help of matrixes for different characteristics this method enables, for instance, manufacturing and market analysis.

Nowadays, in the field of preventive techniques for failure prevention in technological processes, the schematically compounded methods shown in Fig. 8.3 are mostly applied.

In current quality management product-related customer wishes are the sources of motivation for development from the designing process through the manufacturing process up to the delivery of the products.

With the help of *quality function deployment (QFD)* the *voice of the company* can be developed from the *voice of the customers*. The QFD method systematizes this process under the application of matrixes based on the following four steps:

- Customer wishes in terms of product characteristics
- Product characteristics in terms of part characteristics
- Part characteristics in terms of manufacturing regulations
- Manufacturing regulations in terms of production instructions

Every phase can be described by matrixes in the form of a so-called *house of quality*.

This method offers the possibility to affect the production aim in the conception phase and at the same time to obtain information about the critical product and process characteristics for the fulfilment of the customers expectations. Besides the implementation of marketing information in the product, target conflicts between the individual product characteristics may also become visible.

For the detection of potential failure modes during product development, the introduction of new manufacturing methods, and the modification of manufacturing technologies *failure mode and effect analysis* (FMEA) is used. FMEA is especially used in the case of cost-intensive and risk-affected products and processes. FMEA has universal application and is not connected with a special field. At the base of a standardized procedure, which can be supported by corresponding blank forms, the main steps of a FMEA can be divided into risk analysis, risk assessment, the determination of measures, and the evaluation of effectiveness. The risk evaluation results from evaluation of the probability of occurrence, its importance (for the customers), and the probability of detection of the corresponding failure before delivery to the customer. The advantages of FMEA above all are decreased numbers of failures in the early phases of product manufacturing, and in product planning.

The systematic search for imaginable reasons for a failure, called an unwanted event, is possible with the method of *failure tree analysis*. This method, which originated from the field of safety engineering, enables an evaluation of fixed correlations by the determination of the quantitative probability of the appearance of failures.

For this purpose the function of single components (devices) is described under different conditions using a so-called components tree. A subsequent system analysis aims to describe holistically their organization and the behavior of the technical system. The contribution of the individual components to the protection of the overall function of the system, the evaluation of the consequences of the environmental influences of the overall system, and the description of the reaction of the overall system to failures within the system, of resources, and by faulty operations can be described by a failure tree analysis and be calculated or simulated by various evaluation methods.

The methods of *statistical research planning* have the general aim to adjust the relevant product and process parameters using a systematically procedure in such a way that the quality-relevant characteristics closely approach the ideal value with as few experiments as possible.

The weighting of the influencing factors and the quantification of their effect can be solved based on classical statistical research planning using mathematical models (such as factorial research plans); if there are a very large number of influencing factors this can be solved with the help of the empirical procedures developed by Taguchi or Shainin.

The *Poka Yoke* method (from the Japanese: the avoidance of unintentional errors) is dedicated to preventive avoidance of failures in manual manufacturing

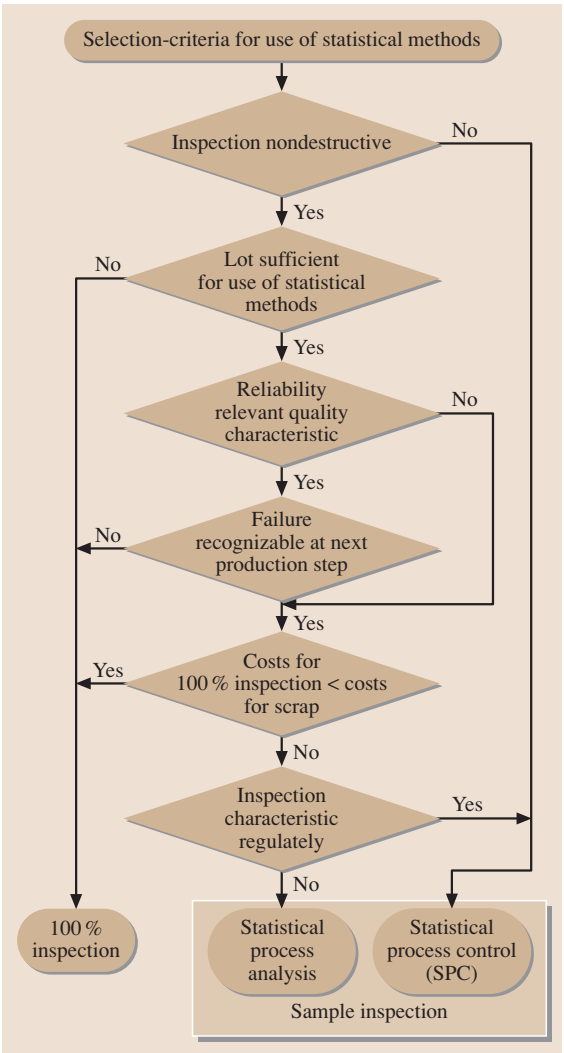


Fig. 8.4 Selection criterions for the use of statistical methods

and assembling processes by the development of precautions (design measures that eliminate incorrect handling) and facilities for failure avoidance for immediate failure detection in the manufacturing process. This can be realized by a comprehensive implementation of applicable rules for product and process design or by the use of simple ancillary equipment.

The *statistical evaluation and control of the quality of processes* [statistical process control (SPC)] is not only a main goal of modern quality management systems but is also required for cost-efficient production processes.

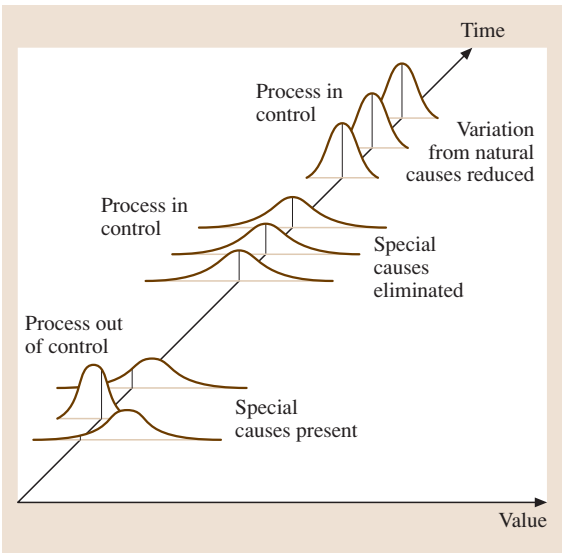


Fig. 8.5 Stabilization of processes through the reduction of systematic and random processes disturbing variables

The aim of the use of statistical methods in quality management exists as well as in the classical detection of faulty products in the manufacturing process and increasingly widespread in the qualitative monitoring, control and improvement of the manufacturing and assembling process.

According to algorithms such as that in Fig. 8.4 some basic decisions can be made.

Quality control of processes can be performed using continuous, 100% inspection or on the basis of periodically chosen samples in the form of statistical process control (SPC).

Statistical methods can also be used to achieve the aim of monitoring and improving product and process quality. Therefore, it is a direct component of the quality control circle. The most important methods in this regard are (Fig. 8.6):

- Measuring controls with statistical control algorithms
- Quality control charts
- Capability analyses and evaluations
- Sample inspection by statistical sample planes

Processes in practice, especially newly introduced processes, are influenced by numerous systematic and random influences, which make the use of statistical methods difficult.

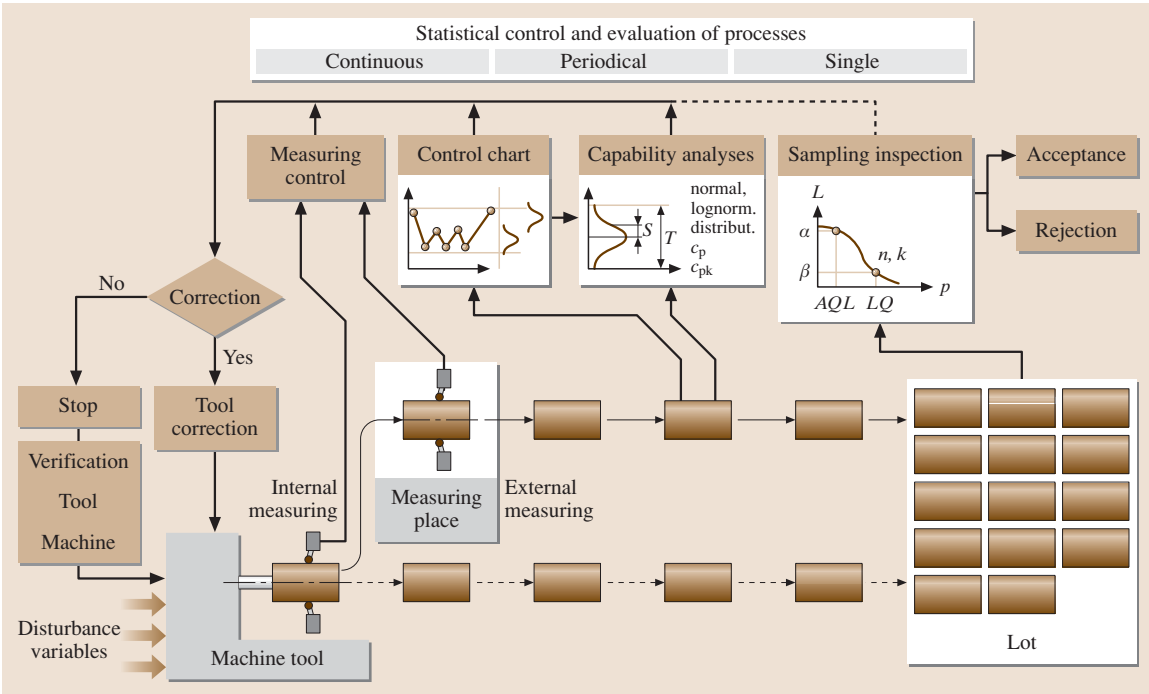


Fig. 8.6 Overview of the use of statistical methods for the evaluation and control of processes

Many statistical methods require the prerequisite of a normal distribution (e.g., double-sided tolerances or limited quality characteristics) or a logarithmic normal distribution (e.g., one-sided toleranced quality characteristics), which are free of systematically influences.

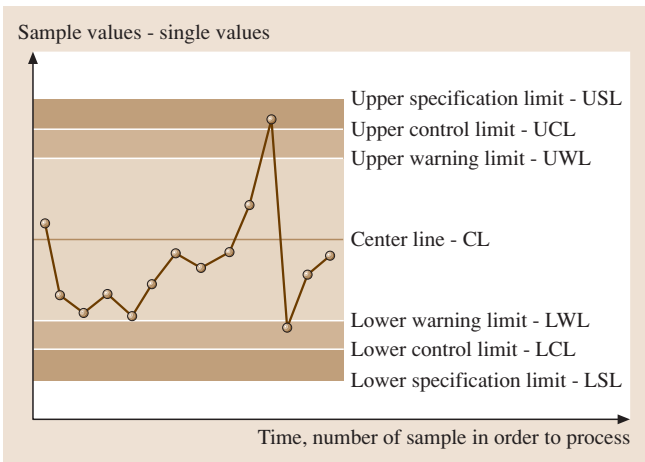


Fig. 8.7 Process monitoring with quality control charts

In such cases it is also spoken from the requirement of stable or stationary processes respectively of so-called *processes in or under statistical control*. (See also Fig. 8.5).

If stable processes are available they can be further inspected using quality control charts. A quality control chart (QCC) is a graphical representation of a process calculated from measured results from a small samples or chronological characteristics. The target of quality control charts is to capture quantitative changes in a process that exceed statistically derived control limits (Fig. 8.7).

The recorded results can then be the basis for subsequent statistical evaluations or characterizations in the form of capability coefficients. These coefficients can also be used for the evaluation of machines and processes.

If a product-orientated decision about the acceptance or rejection of a product series or lot takes place, sample planes can be used. The producer (supplier) may also use these in the form of a final inspection, as may the customer in the form of an inspection on receipt.

Statistical analysis results have a wide variety of uses in quality monitoring, evaluation, control, and management as is shown in Fig. 8.8.

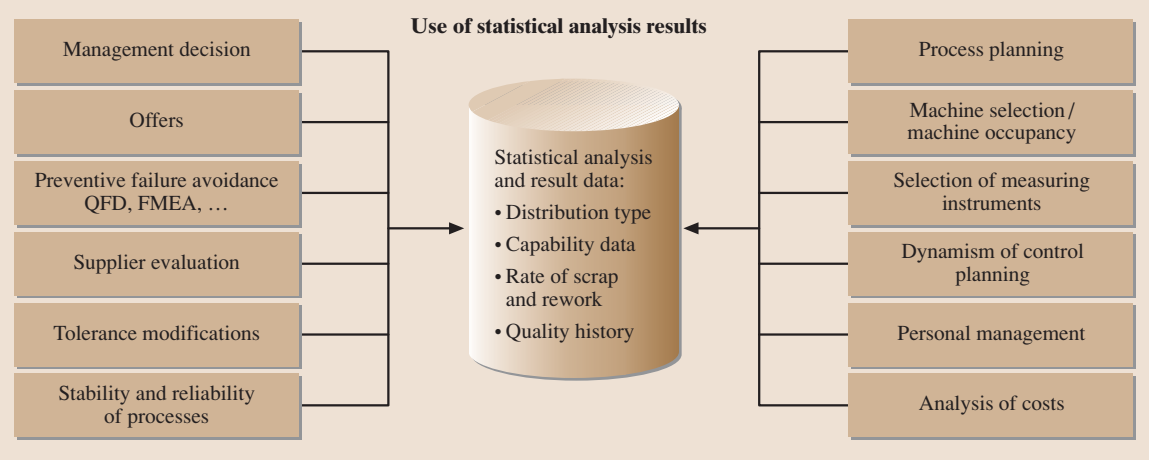


Fig. 8.8 Use of statistical results

8.1.3 Quality Management Systems

The organizational realization of the concept of quality in all departments of enterprise and in all phases of product realization can be supported by *quality management systems* (QMS), as shown in Fig. 8.1. In addition to this, the organizational structure and the processes organization of all quality-relevant departments and operations can be evaluated and compared to standardized regulations. The successful introduction and realization of a QMS can be evaluated and validated by an accredited certification body.

As the basis for this the independent ISO 9000 series of standards can be used. Based on this, for the worldwide automotive industry, comprehensive additional requirements exist, which are fixed in company-dependent standards. Some of the most important series of standards are, for instance, QS 9000 (USA), VDA 6 (Germany), AVSQ (Italy), and EAQF (France). Efforts for the integration of the different requirements of the automotive industry to date are contained in the ISO standard TS 16949.

8.1.4 CE Sign

Within the single market of the European Union, a requirement for the commencement of operation and the placing on the market of products is the fulfilment of harmonized safety requirements. To obtain a conformity certificate for these European Community (EC) requirements, a CE sign is required. This is also valid for products that are produced outside of the European Union, if they are also distributed within it.

Examples of EC directives include those for: machines, toys, electromagnetic compatibility (EMC), detonating devices, and medical products etc.

8.2 Manufacturing Measurement Technology

8.2.1 Introduction

Measurement is the determination of the value of a physical quantity (measured quantity) by comparison with a reference value.

Manufacturing measurement technology is part of the field of measurement technology. It deals with the methods, equipment, and strategies for measurement in the realm of mechanical production processes. Manufacturing measurement technology is utilized in various locations, such as rooms with special, controlled environmental conditions, as well as directly integrated into the assembly line. The aims of manufacturing measurement technology are:

- Evaluation of the product
- Evaluation of processes and machines
- Quality-orientated process control

Here we confine the discussion to the description of relevant geometrical features and/or technical aspects of obtaining such objects from workpieces. These features include lengths, angles, distances, and surface struc-

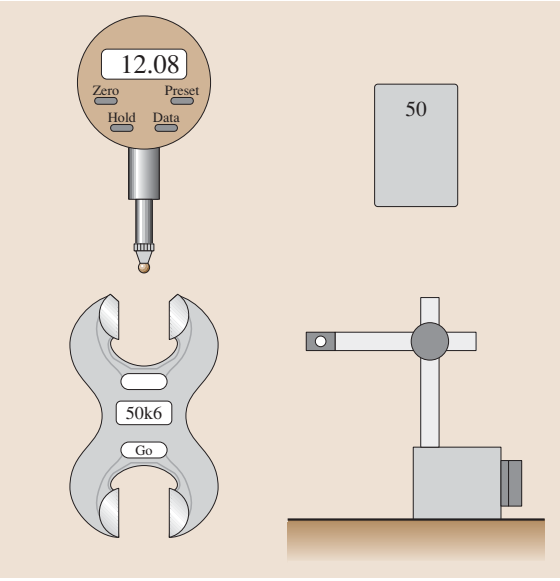


Fig. 8.9 Inspection devices

tures. Most of the principles and definitions described also apply to nongeometrical parameters.

The relevant SI units used to describe these geometrical features are meter and second. Angles are described in degree or radian.

The physical measurement that comes from the measuring process is called the measure. The object from which this quantity was measured is called the measuring object. The value of the measure, which actually exists but is currently unknown, is the true value. The aim of the measuring process is to determine this true value. A variable that does not change with time (for example, a diameter) is measured by a static measurement. Dynamic measurement is the measuring of a variable that changes with time (for example, a vibration) or the measurement of a variable, whose variation arises from the time-dependent behavior of another variable (for example, for a roughness measurement, the dependence of the measured value on the scanned measuring length).

In order to determine whether an object fulfils a particular requirement, it is inspected. As subjective inspection should not be considered, one can differentiate objective inspection into measuring and gauging. The inspection means that are employed can be divided into indicating measuring devices, measurement standards, gauges, and additional measuring devices. Figure 8.9 shows a digital dial indicator and a measuring block, as the most important measurement standard of dimensional measuring technology. Measuring blocks are quadrilateral metal or ceramic gauges, whose parallel end faces have a known width and minimal deviation. A gauge and a stand, an example of additional measuring equipment, are represented on the figure to.

8.2.2 Arrangement in the Manufacturing Process

Measuring technology has become an essential ingredient in manufacturing processes. Figure 8.10 shows the multitude of integration possibilities and the corresponding relationships between the production process

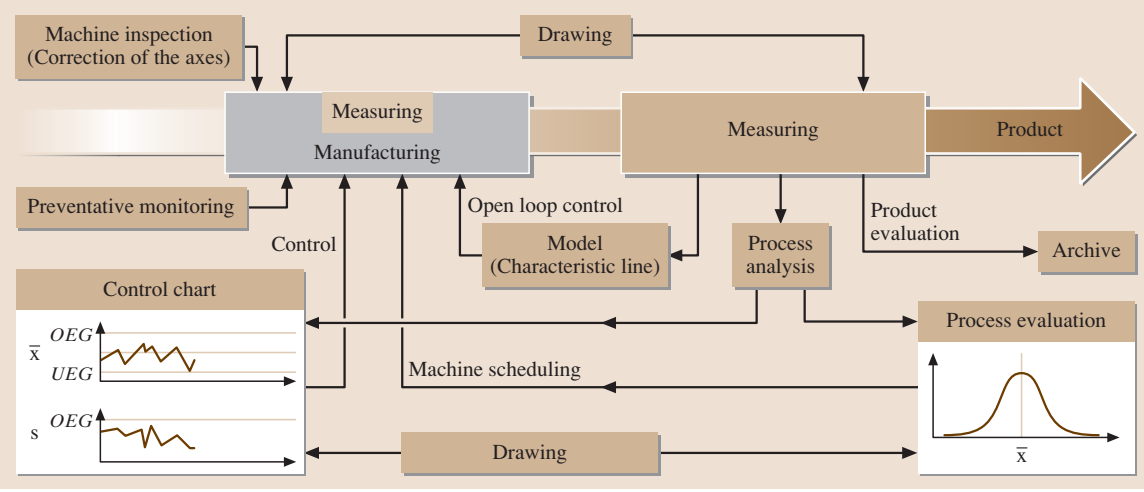


Fig. 8.10 Measuring technology and measuring value analysis in manufacturing processes

and measuring technology as well as different possibilities for analyzing the measured data and its feedback into the process.

Inside every tooling machine integrated measuring devices are used to collect parameters (the position of the axes, temperature, force, voltage, current, etc.). These values can then be directly used by the machine operator or the machine itself to control future operations. In reference to the collection of data from finished products and in consideration of the drawing specifications, these finished product parameters can be evaluated, classified, and documented. Furthermore, parameters can be derived from the measured values through continued analysis, or more specifically, these parameters that can regulate or control the machining or production of the product. Using an adjustment, the measured value (the controlled variable), which has an instantaneous value, is continuously measured and corrected with a controller with respect to a known value (the set value). It is then possible to eliminate disturbance factors from outside the process. The constant, targeted influencing of a process on the basis of a previously defined model without feedback is what is meant by open-loop control.

Starting from the process analysis, the machining process and/or the processing machines themselves can be evaluated (for example, for machine and process capabilities) and documented. The result of the machine evaluation is, for example, a quality-orientated machine selection. Should the machine not be in a position to produce the desired tolerances, selective inspection and correction of the machine systems may be necessary, or a complete overhaul or a new revision of the machine may even be necessary.

Through the additional use of measuring technology for preventive monitoring of a machine, it is possible to implement maintenance procedures or schedules to avoid losses or stops in production.

8.2.3 Specifications on the Drawing

The drawing is the defined input to the manufacturing process, based on the required functions outlined by the designer, and is therefore the basis for the manufacturing measuring process. The complete, integrated product drawings, as well as single-product element drawings, are used to clarify the permitted allowances and thereby the tolerances for manufacturing. The geometrical specifications can be separated into measuring tolerances and shape tolerances. Shape tolerances include not only form and position tolerances, but also surface finish tolerances, which can be divided into roughness and waviness.

Measuring specifications consist of the nominal size and defined permissible upper and lower allowances. The nominal size plus the upper allowance is the largest acceptable measure and the nominal size plus the lower allowance is the smallest acceptable measure. The difference between the upper and lower allowances is the tolerance. The definition on the drawing is shown through the specification of a nominal size supplemented either directly by a coded specification of the allowance or through a letter and a number combination (ISO 286). The letter describes the position of the tolerance field with respect to the zero line and the number represents the magnitude of the tolerance. The specification $36^{-0.025}_{-0.050}$ corresponds to the drawing specification, 36 f7, which means in manufacturing terms

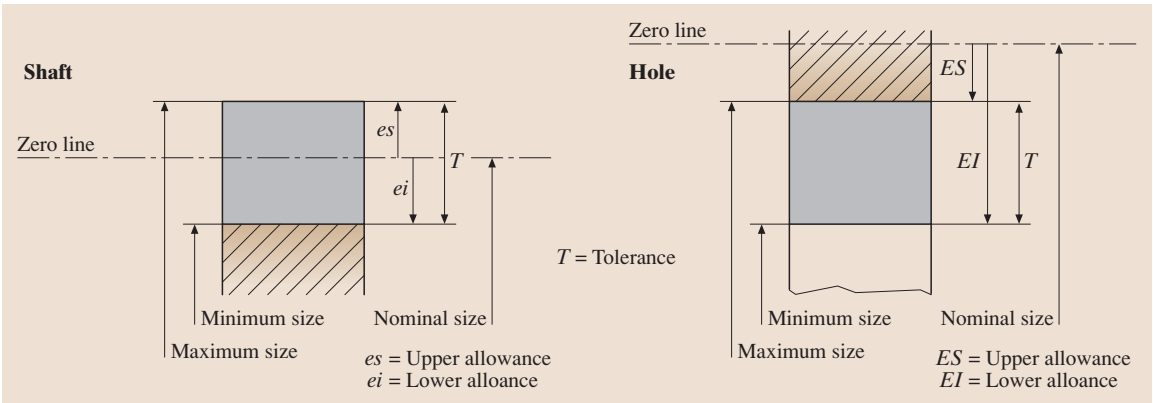


Fig. 8.11 Measuring specifications

Table 8.1 Form tolerances

Form tolerances	Straightness	—	Single elements
	Flatness	▱	
	Roundness	○	
	Cylindricity	⌭	
	Profile	⌒	

that the largest acceptable measurement is 35.975 mm and the smallest acceptable measurement is 35.950 mm. Whenever the nominal sizes are the only values displayed on the drawing, free size tolerances according to ISO 2768, can be specified.

The position of the tolerance with respect to the zero line (which corresponds to the nominal size) and the size of the tolerance determine the function of the proposed piece. The size of the tolerance is furthermore decisive for the cost of the manufacturing process.

Table 8.2 Position tolerances

Position tolerances	Direction	Parallelism	//	Related elements
		Perpendicularity	⊥	
		Angle	∠	
	Place	Alignment	⌖	
		Concentricity	◎	
		Symmetry	⌯	
	Run	Run-out	↗	
		Total run-out	⌰	

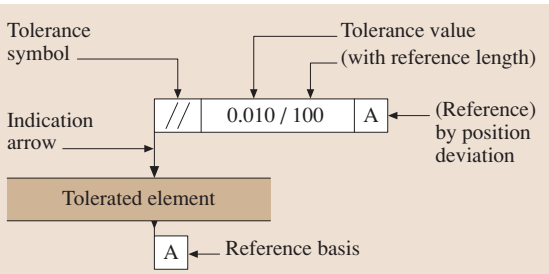


Fig. 8.12 Tolerance frame as drawing entry

The form and location tolerances define in each case zones, into which the corresponding element must fit in, when it is within the tolerance. Form tolerances, which are defined in ISO 1101, are shown in Table 8.1. Single elements, such as lines, rectangles, circles, cylinders, and profiles are specified as form tolerances. Position tolerances are divided into direction, place, and run tolerances, as summarized in Table 8.2. In each case, position tolerances refer to a individual element of the workpiece. On the drawing, the necessary specifications are represented in a tolerance frame (Fig. 8.12). The tolerance frame contains a symbol that describes the type of tolerance that is being signified and a tolerance value, possibly with a reference length. The indication arrow connects the tolerance frame with the associated element, or rather the associated lines of measurement or symmetry on the drawing. Whenever the type of tolerance requires it, the reference element is indicated at the end of the tolerance frame through the specification of a letter, which repeats the necessary reference basis. In a few cases, more reference elements are possible.

The tolerance specification for surface finish is shown on the drawing by a basic symbol and supplemented by additional specifications. Whether material cutting machining is *necessary* or *not permissible* is also indicated, as shown in Fig. 8.13. additional specifications include tolerances for the roughness value [most of the time, the arithmetic average surface finish (Ra)], and

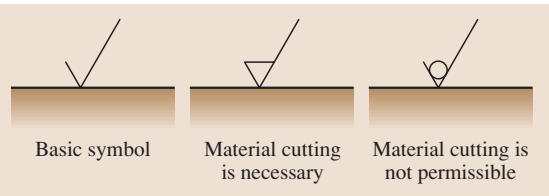


Fig. 8.13 Symbols for surface conditions

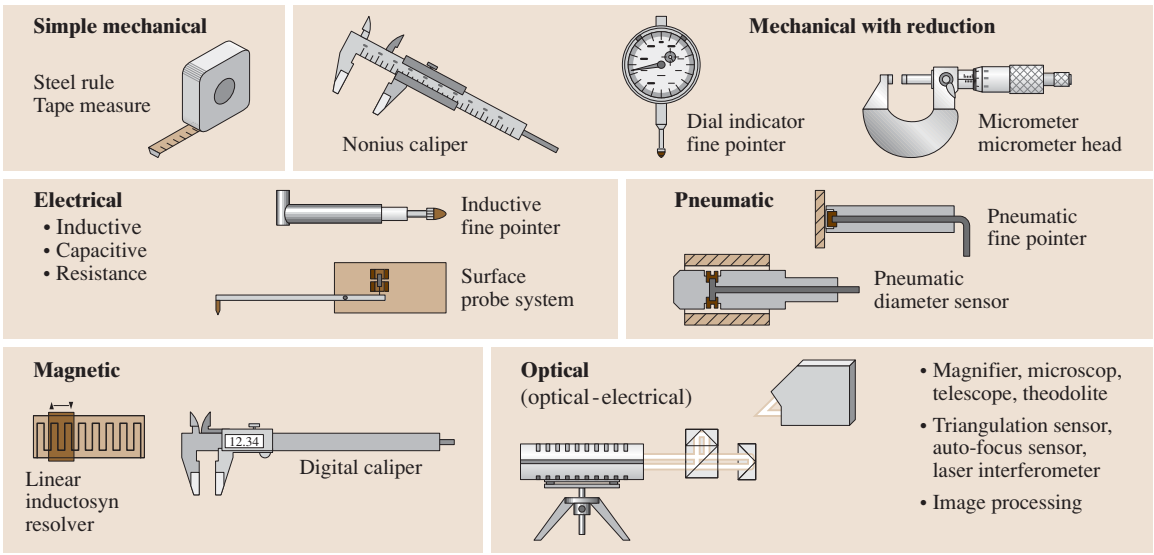


Fig. 8.14 Physical principles of measurement value collection and example instruments

when necessary, amongst others, specifications for the manufacturing process, tolerances for further roughness or waviness parameters, and the grooving direction.

8.2.4 Gauging

Gauges are inspecting devices that represent the dimension and/or form of the workpiece that is to be tested. There are three types of gauges: limit gauges, receiving gauges, and so-called *go and no-go* limit gauges. The procedure is always the same: to try to mate the workpiece elements with the correct gauges. A simple yes or no decision is the result of this process. With limit gauges, one can determine whether the value of a test object is larger or smaller than the value of the gauge (for example, with a pin gauge). With receiving gauges (for example, radius or screw pitch gauges), a comparison is made with the workpiece's form and the form of the gauge and the best-fitting pieces can then be separated out.

Go and no-go limit gauges are the most important for the inspection of manufacturing toleranced elements. Figure 8.9 shows an example of an external calliper gauge, which is suitable for the inspection of shafts. Limit gauges consist of a *go* and a *no-go* side, which correspond, respectively, to the largest and smallest measurement possible; the *go* and *no-go* sides are verified one after the other to check for mating. The inspection result is *good* when the workpiece fits comfortably into the *go* side and not into the *no-go* side. The Taylor principle states that, on the *go* side, all the measurements or control measurements are tested at the same time, but on the *no-go* side, every measurement or control measurement is tested in isolation.

8.2.5 Application of Measuring Devices

Measuring devices work on very different physical principles. Simple mechanical measuring devices include the steel rule or measuring tapes. Measuring devices with mechanical reduction include nonius calipers, dial

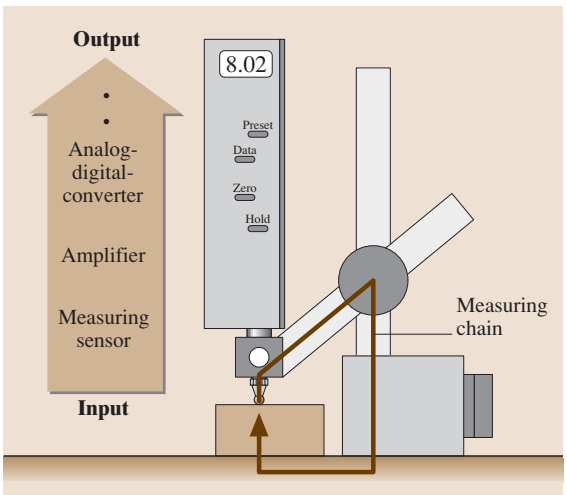


Fig. 8.15 The measuring chain

indicators, and micrometers. Pneumatic systems have long been used because of their noncontact scanning. At present, electrical, magnetic, and optical (more specifically, optical-electrical) principles have emerged in the field of measurement. Their common essential property is that the measured value can be directly transformed into an output signal or onto a recording media. A few examples of equipment are shown in Fig. 8.14.

Essential for a secure measuring value collection is the construction of a measuring chain that is as short as possible, and stiff. The measuring chain in Fig. 8.15 is composed of a measuring device, the supports of a stand, a base plate, and the measured object. In the measuring device, starting from the input value (in this case, the displacement of the probe by the measuring object) the output (in this case, the measuring value 8.02 mm) is obtained through the measuring sensor, an amplifier, and an analog-to-digital converter. Depending on which measuring devices are used, their internal parts may be quite different. The output value can then be transferred to a storage device or a processing machine (for example, a personal computer).

The following explains several essential properties of measuring devices. These properties determine the applicability of the equipment. The achievable *information content* of a measurement is an essential distinction criteria. Also, the *accessibility* of the measuring value to the workpiece reduces the possible selection of measuring devices. The *type of probe* (contact and noncontact) is also relevant. Spheres, planes, or cutting edges are the fundamental probe shapes.

The *indicating range* of a measuring device is the range of values over which the device can indicate (smallest to largest indicated measuring value). The *measuring range* can deviate from the indicating range, and corresponds to the section of the indicating range over which the measuring deviations remain within defined boundaries. The *scale division value* is the change in the measuring value that leads to the movement of the pointer from one to the next dash on a line scale (on digit scales to a digit step). When the measuring devices are clamped (for example, by a support Fig. 8.15), the range of adjustment of the support and the measuring range of the measuring devices constitute the *range of application*. The relationship between the value of the input and the corresponding output value is the *transfer function*. A linear transfer function is usually aimed for. The *threshold limit* is the smallest possible change in the input that leads to a recognizable change in the output. The effect whereby the same value of the input produces two different results

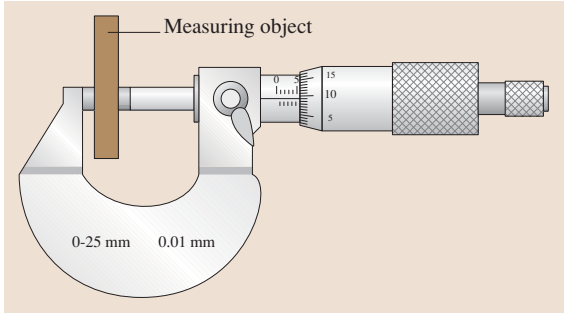


Fig. 8.16 Direct measuring

depending on the direction in which the value is approached, is called the *hysteresis*. A slow, time-varying change of the output without a change in the input value is called *drift*. The *response time* is the time between a step change of the input and the time when the output value remains constant within a prescribed range.

Additionally the *measuring uncertainty* should be highlighted as an essential characteristic of measuring devices. A measuring device is qualified for a designated measuring task only when the uncertainty does not exceed a ratio with the tolerance of $\frac{u}{T} = 0.1$ to 0.2. Because the exact uncertainty of a chosen measuring devices is often not known, one can make do with the *limit of error*. The limit of error is the largest measuring deviation of a measuring device, and is usually defined for a group of measuring devices through a standard or by the manufacturer.

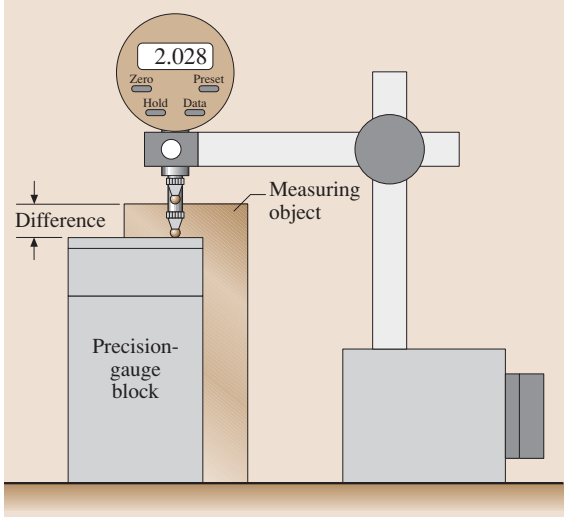


Fig. 8.17 Difference measuring

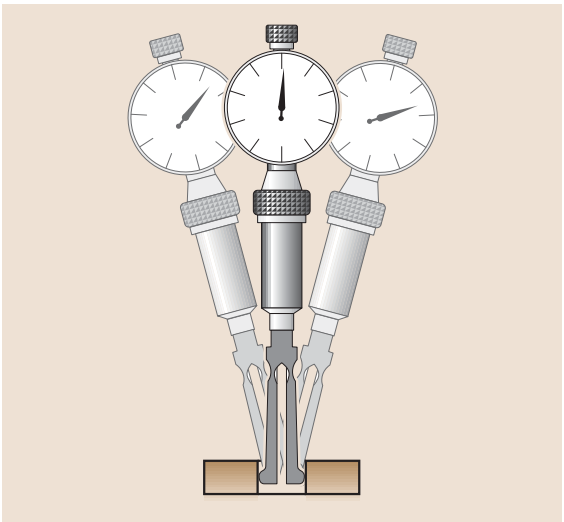


Fig. 8.18 Inversion point

The *direct collection of measuring values* (for example, with callipers or micrometers) is the easiest type of measuring implementation. The measuring range of the measuring devices must overlap with the measuring object. In Fig. 8.16 the measurement of a length by means of a micrometer is shown. The range of measurement of the displaced micrometer is shown as being 25 mm. Objects with lengths ranging from 0 to 25 mm can be measured. The measuring uncertainty for such a micrometer can be expected to be in the range 10–20 µm.

If the tolerance of the measuring value is so small that the required ratio between the measuring uncertainty and tolerance can no longer be met, one must choose a measuring devices with a smaller measuring uncertainty. Smaller measuring uncertainties are often accompanied by smaller measuring ranges. Whenever, for example because of the requirements of the measuring uncertainty, a measuring device whose range of measurement is smaller than the measuring value must be used, one can employ *difference measuring* (Fig. 8.17). One chooses a measurement standard, in the figure shown as a combination of block gauges, whose value differs from the measuring value by no more than the measuring range of the measuring device. The measuring device is used so that the readout both by probing the measurement standard and the measuring object is possible. The length of the measuring object is the sum of the length of the precision gauge block and the difference between the readout for the measurement standard and the measured object.

Another important application option is the search for an inversion point. Figure 8.18 shows an example of the measurement of an inside diameter with a two-point measuring instrument (two probing points on the cylinder's interior). Through the measuring force and supported by special structures, the instrument aligns itself so that the cylinder axis intersects the connecting line between the two probe points. The instrument can be pivoted from the left in Fig. 8.18, through the middle, to the right position. One can see that the displayed measuring value at first becomes smaller but, after the middle position, becomes larger. The smallest displayed value, the inversion point of the needle, corresponds to the shortest connection of the probed cylinder surface line, and thus the diameter.

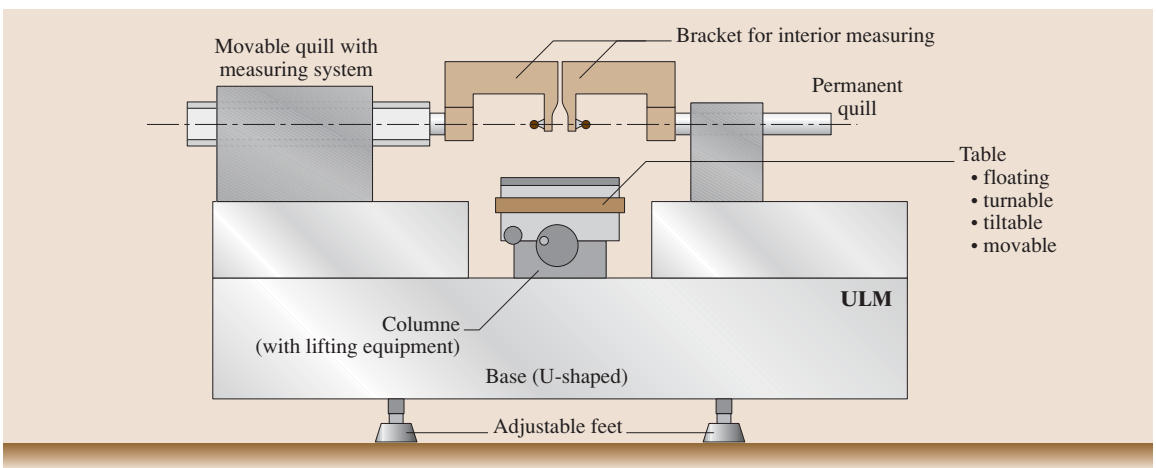


Fig. 8.19 Universal measuring gauge

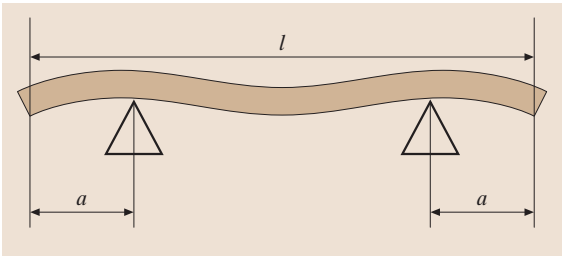


Fig. 8.20 Bessel points

When measuring workpieces, different influences have a negative impact on the measuring result. For example, inclination between the measuring instrument and measured object, bending, deformation on account of the measuring force, and environmental conditions can lead to deviations of the measurement. Some of these effects can be avoided or notably reduced through proper equipment operation.

Through the adherence to a fundamental measuring technology principle, the Abbe comparator principle, one can reduce the influence of tilting errors. According to this principle, the measured object and the measuring rule are to be aligned flush. Figure 8.19 shows the universal length gauge as one of the essential instruments using this principle. The measured object and the integrated measuring system, which is inside of the movable quill, are consecutively aligned flush. Brackets for interior measurement are attached. After removing the brackets the instrument is suitable for the collection of external measurements. As measuring systems, glass measuring scales and laser measuring systems are utilized. Such measuring instruments are suitable for the monitoring of gauges and measuring means.

The deflection of workpieces due to their own weight also has a considerable impact on their length. On a bar-type component, one obtains the smallest shortening of the whole length whenever the piece is supported on the so-called Bessel points (Fig. 8.20). These support points are located at a distance $a = 0.220 \cdot l$ from the ends. Similar effects lead to deformations on account of the measuring force. Deformations of the material are mostly only problematic for soft materials (plastics) and non-massive workpieces. Here, the workpiece can bend itself whenever the measuring force presses against thin-walled positions. This effect can be reduced using proper fixtures.

Environmental conditions are generally specified by the room where the measurement takes place. Measuring rooms are classified into five classes. Precision measuring rooms (class 1) serve for the calibration of reference standards. Fine measuring rooms (class 2) serve for the calibration of used standards and for acceptance inspection of precision parts. In standard measuring rooms (class 3), measuring tasks for the monitoring of a process, measuring of fixtures and tools as well as the control of inspection equipment (factory standards) and first prototype inspections are conducted. The production-related measuring room (class 4) serves for monitoring production, machine settings, and instruments. Production-related measuring takes place on the manufacturing measuring site (class 5). The spectrum is supplemented by the special measuring room, for example, for the measuring of semiconductor wafers. Properties whose limit values in every measuring room class are predefined include temperature (basic temperature, time and local temperature variation), vibrations, air quality (fine dust, suspended matter), air humidity, and lighting.

Table 8.3 Linear expansion coefficients

Material	Linear expansion coefficient (µm/(m K))
Steel	13
Copper	16
Aluminum	23
Magnesium	26
Polyvinylchloride	80

A substantial influential parameter for geometrical data acquisition is temperature. The reference temperature for measurement is 20 °C. The reference temperature is the temperature at which workpieces have their true measurement value and at which the inspection equipment detects it. These specifications apply to all pieces in the measuring chain. Table 8.3 shows the linear expansion coefficients of different materials. For example, for a temperature increase of 1 K, a 1 m-long piece of steel will become about 13 µm longer. The alternative measuring of the existing temperature and consequent correction of the measuring value is particularly problematic for geometrically and compositionally inhomogeneous inspection equipment or workpieces and by temporally and spatially varying temperatures.

8.2.6 Coordinate Measurements

Basics

Coordinate measuring machines are universal, flexibly applicable equipment for the recording of workpiece geometry. As well as measurements of parameter (for

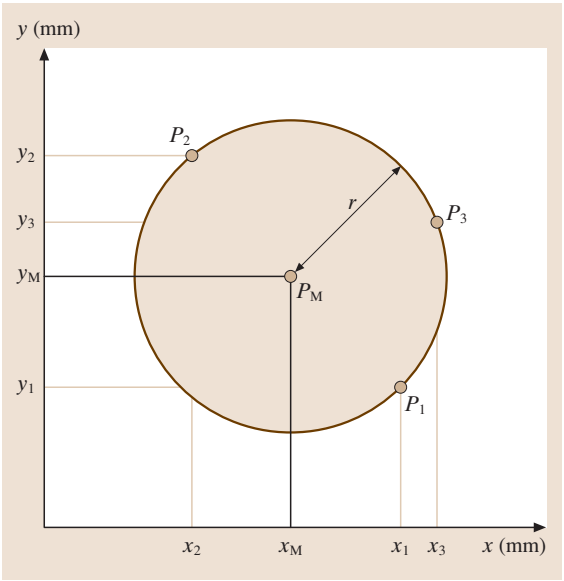


Fig. 8.21 Measuring points on a circle

example, diameter and length) position and form (for example, roundness and concentricity), specials (for example, cylindrical or bevel gears) and free-formed surfaces (castings) can be measured.

In contrast to conventional measuring techniques the desired measurements are not measured directly. The principle is the acquisition of single data points from geometrical elements. These elements include circles, planes, cylinders, spheres, and cones. Furthermore, these elements can be combined together to identify distances, angles, or position deviations. The acquired measuring points, which are on the boundary of standard elements, are mostly described by a Cartesian coordinate system (x, y, z). However, it should be noted that cylindrical and spherical coordinate systems are also possible. The described size of these geometrical elements will then be calculated from these measured points. The following example shows the strategy for identifying the radius r and the middle point coordinates x_M, y_M of a circle. For simplification, this example will be confined to a circle in the x–y plane.

The equation for a circle whose center point does not lie at the origin is as follows

$$r^2 = (x - x_M)^2 + (y - y_M)^2 \,.$$

The variables r, x_M and y_M are unknown. One appoints a measuring point on the circumference of the circle for the variables x and y. Now, because there are three unknowns and we need a clear description of the circle, we need three equations, which means three data points (the smallest number of data points for determining a circle

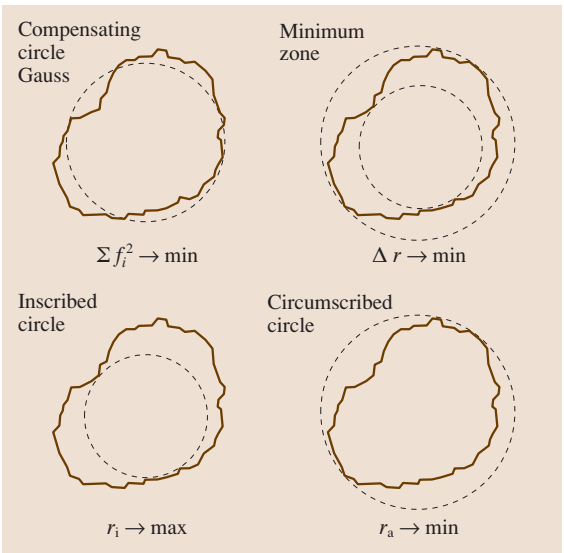


Fig. 8.22 Compensation methods

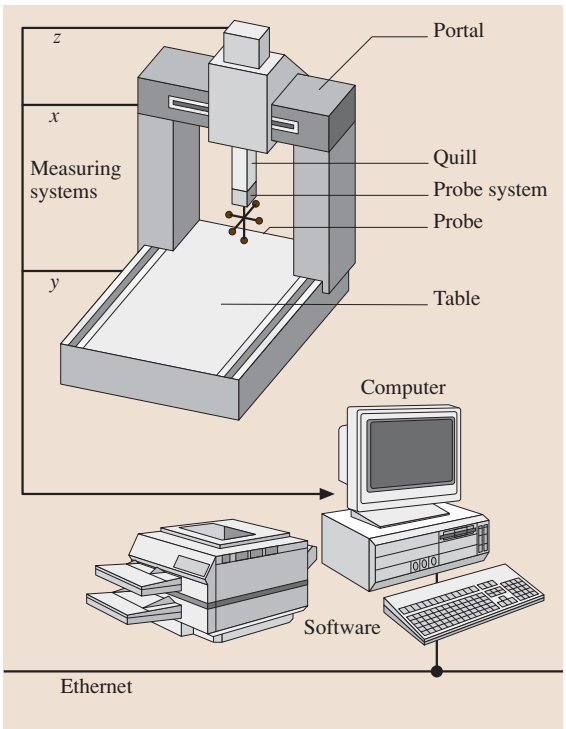


Fig. 8.23 Basic configuration

is three). This yields the following system of equations

$$\text{For } P_1 \text{ with } [x_1, y_1]: \; r^2 = (x_1 - x_M)^2 + (y_1 - y_M)^2 ,$$
$$\text{For } P_2 \text{ with } [x_2, y_2]: \; r^2 = (x_2 - x_M)^2 + (y_2 - y_M)^2 ,$$
$$\text{For } P_3 \text{ with } [x_3, y_3]: \; r^2 = (x_3 - x_M)^2 + (y_3 - y_M)^2 .$$

Since we are dealing with elements of measured, real-world objects, whose bordering areas deviate from geometrical standards, the collection of a large number of data points is essential for the determination of the elements parameters. An overdetermined (no longer explicitly solvable) system of equations results from a higher number of measured points. This system has the following form

$$\text{For } P_1 \text{ with } [x_1, y_1]:$$
$$f_1 = (x_1 - x_M)^2 + (y_1 - y_M)^2 - r^2 ,$$
$$\text{For } P_2 \text{ with } [x_2, y_2]:$$
$$f_2 = (x_2 - x_M)^2 + (y_2 - y_M)^2 - r^2 ,$$
$$\vdots$$
$$\text{For } P_n \text{ with } [x_n, y_n]:$$
$$f_n = (x_n - x_M)^2 + (y_n - y_M)^2 - r^2 ,$$

where f_i represents the deviation of the corresponding data point from an ideal circle. One can employ a compensation method to solve this system of equations. A widely used method is the regression equation by Gauss. The sums of the squared deviations are minimized with this method

$$\sum_{i=1}^{n} f_i^2 \rightarrow \min .$$

This least-squares method ensures that all measuring points are included in the calculation. Problems especially arise when one wishes to verify mating. In this case the inscribed circle gives the description for holes, while the circumscribed circle gives the description for shafts. (The *inscribed circle* is the largest possible circle that has all of the measuring points outside, while the *circumscribed circle* is the smallest possible circle that contains all of the measuring points.) In certain cases an outlier test may be necessary.

Equipment, Sensor Technology, and Software

One can subdivide coordinate measuring machines on the basis of their construction into portal, bridge, and standing measuring machines. In Fig. 8.23, the basic configuration of a portal measuring machine is repre-

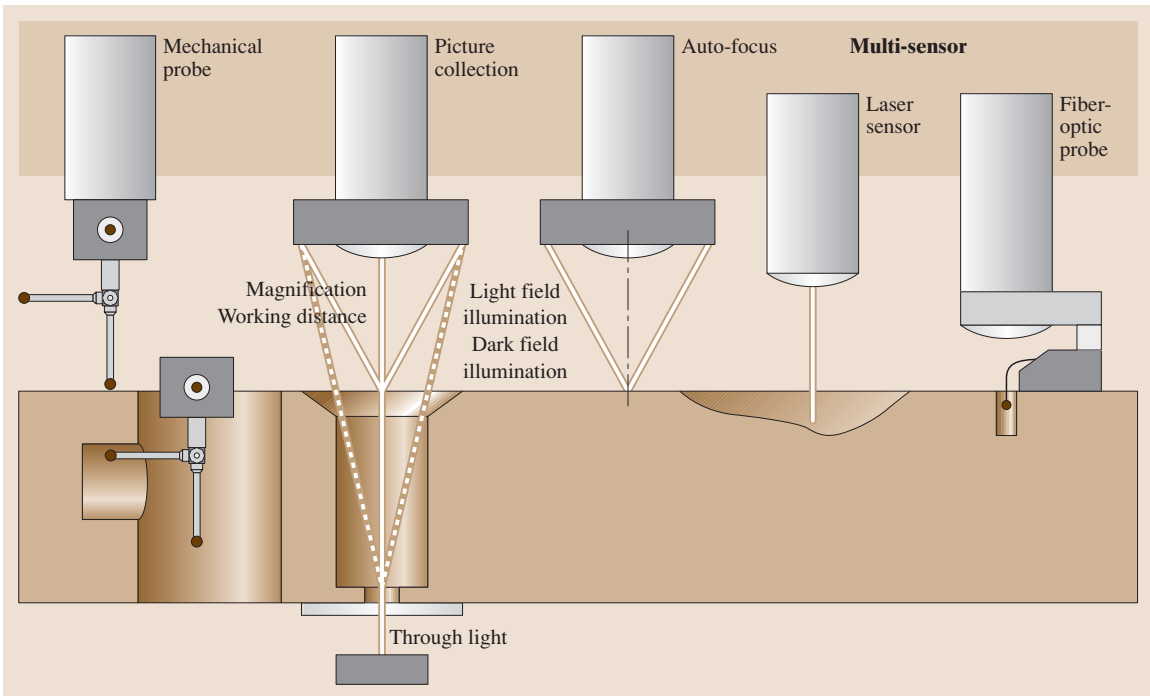


Fig. 8.24 Multisensor technology

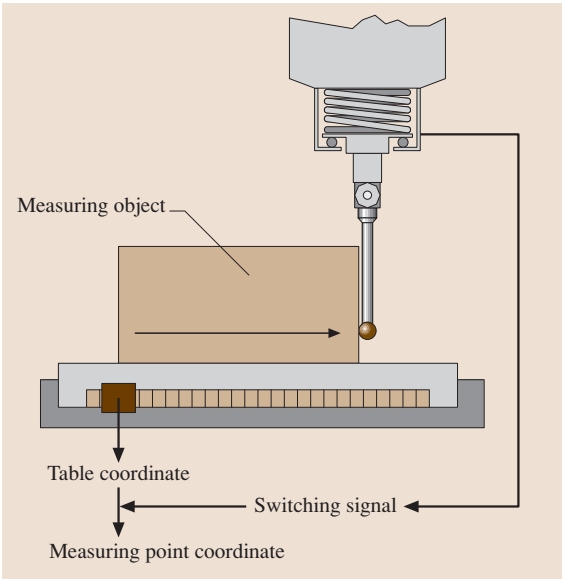


Fig. 8.25 A switching probe system

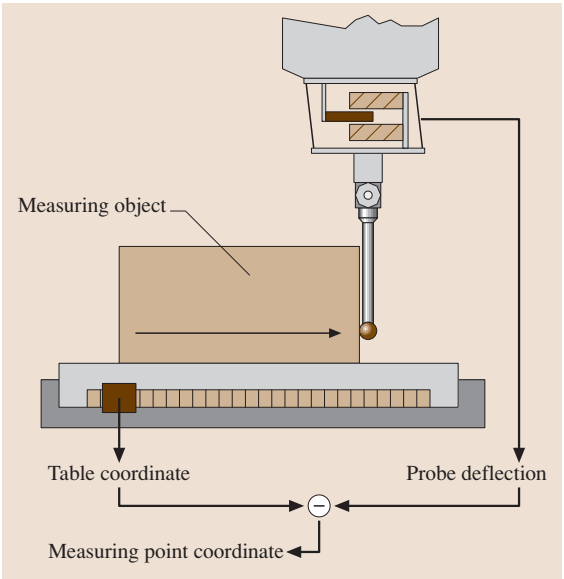


Fig. 8.26 A measuring probe system

sented. The mechanically embodied axes (in the shown example, portal and quill) are arranged on the measuring table at right angles to each other. Additionally, the axes of rotation can be integrated for the realization of rotary movement. The instantaneous positions of the axes are logged through the measuring system and forwarded to the connected computer. Glass measuring scales are mostly used in measuring systems.

The probe system is attached to the quill. In addition to the frequently utilized, mechanical probe systems there are contactless options such as edge recognition, optical methods, and/or laser probes. One speaks of a multisensor coordinate measuring machine when various probing systems are arranged inside (Fig. 8.24).

Figures 8.25 and 8.26 illustrate various mechanic probing systems. One can distinguish between measuring and switching mechanical probing systems. In both cases, the workpiece is touched with the probe element, which is fixed onto the probing system. The probe element consists mostly of a shaft and an almost ideally round, wear-resistant sphere (in most cases, made of ruby). In a switching measuring system (Fig. 8.25), the contact between the probe sphere and workpiece releases a impulse. At the moment of this signal, the x, y, and z coordinates of the current position from the table, portal, and pinole are transferred to the computer and stored there for further processing as required.

Contact between the probe and workpiece leads to deflections of the probe head in all three coordinates in a measuring probe system (in Fig. 8.26, the deflection is only represented in one direction). These deflections are measured and combined with the associated coordinates of the inherent positions of the table, portal, and quill. The measuring point coordinates obtained in this way are passed to the computer. However, care must be taken to note that the measuring point coordinates acquired always refer to the middle point of the probe sphere of the reference probe, not to the contact point between actual probe and workpiece.

Very flat outlines or color transitions are not mechanically probable. In these cases, optical acquisition is an option. For this purpose, the measuring object is illuminated, according to the target measurement (using through light, light-field or dark-field illumination) and imaged through optics. The optics used may work with changed or zoom objectives in order to obtain different magnifications between the measured object and the image. In the simplest case, only the contours to be measured are scanned, using an edge finder. In the next most sophisticated system, based on image processing, the measured object is displayed on a charge-coupled device (CCD) matrix. As by a mechanical measuring probe system the measuring point coordinates results as a combination of the pixel coordinates and the inherent positions of the table, portal, and pinole.

While the optical variants described so far operate in the x–y plane, the focusing of an optical system can also capture the z-coordinate. The focusing can be auto-

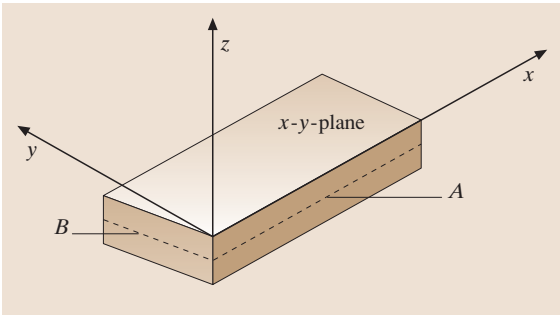


Fig. 8.27 The workpiece coordinate system

mated by means of a laser or contrast evaluation in the image-processing system. The sensor controls the shift of the quill in the z-direction. The respective z-value supplements the x–y coordinates of the associated position to actual measured three-dimensional (3-D) values. Instead of autofocusing options, one can also use a direct laser measuring system for the addition of the third coordinate, for example, a triangulation sensor. This method forms a laser point on the surface of the workpiece. Using observational optics arranged at a defined angle with respect to the sensor, a difference in height of the surface of the workpiece is recorded directly. When the measuring range of the triangulation sensor is not sufficient, one can extend its application range by shifting with the quill. The measured z value can then be calculated from the z-axis of the measuring system and the measured value from the triangulation sensor.

One very new and interesting sensor variation is the fibre-optic sensor patented by *Werth-Messtechnik* (see also Fig. 8.24). It consists of an extremely small glass sphere, which is suspended by a glass fibre. Over this light pipe, the sphere can be directly illuminated. To capture the measuring points, the glass sphere is placed on the workpiece to be measured. The image-capturing system (CCD matrix) can detect this probe either in transmitted light or through its self-lighting. The center of the sphere describes the coordinate of the measuring point, just as for normal mechanical probes.

The *software* in the attached computer essentially has the following tasks:

- Capture of the measuring points
- Corrections, for example, perpendicularity and guideway deviations of the measuring machine components, probe radius and middle point, probe bending, temperature influences
- Coordinate transformations, for example turntable and workpiece coordinate systems
- The calculation of ideally geometric substitution elements from the measuring points
- The combination of single elements
- Conversions, projections, etc.
- Evaluation (nominal-actual comparison), protocol, statistics

Device control (point- or path-control, scanning) comes with the computer numerical controled (CNC) coordinate measuring machines.

For the archiving of results or for further processing data transfer by means of a network is possible.

Proceeding with the Measurement

The completion of a measuring task by means of coordinate measuring machines is explained in the following. Three phases will be differentiated.

Phase 1: Planning. Planning can take place far from the machine but is based on knowledge of the technical capabilities, available probes and clamping elements as well as extensive experience. This phase is decisive for measuring-technically proof and economically justifiability of the measuring process. The test task, which is generally fixed on the drawing, forms the basis for the planning of the measurement. Differences in the path arise, whenever the aims of measurement (evaluation of the part), quality-orientated production control or check for assembling possibility are changing. In this case knowledge about the production of the piece (for example, the machining base) or its later application (subsequent machining and assembly) is useful. The test task is divided into its single elements and combinations. A sequence from easiest to most complicated is then sought.

It is to ones advantage to specify a coordinate system on the workpiece not only for the measurement, but also for the evaluation of the results. This workpiece coordinate system must be mathematically clearly described. To do this, the following steps are required:

1. Space adjustment: description of a coordinate axis (mainly z) as the main direction of important form elements (for example, the direction normal to a surface, the axle of a cylinder). The z-axis is described in the example picture (Fig. 8.27) by the normal to the upper workpiece plane.
2. Plane adjustment: hindering the rotation of a coordinate system around these under 1. defined axes (for example, by fixing to an edge or the straight line connecting the middle of two holes). The direction of the x-axis in the example picture (Fig. 8.27) is

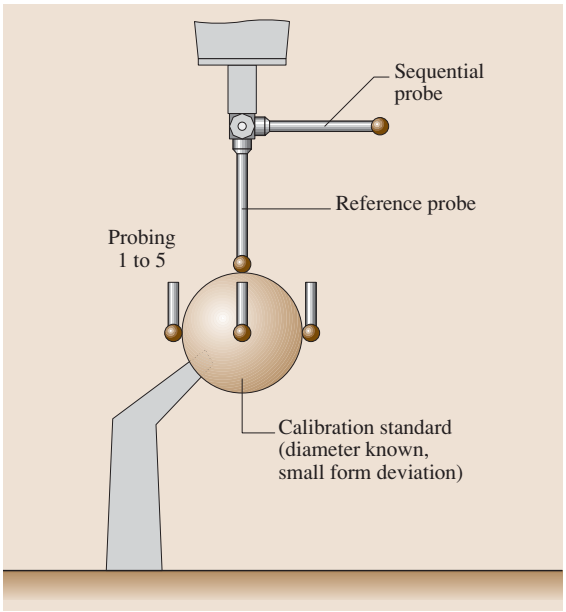


Fig. 8.28 Probe set-up and calibration

defined in the x–y plane by the corresponding direction of the long rectangular surface (dashed line A). This fixes also the direction of the y-axis, which must be orientated at a right angle to the x-axis.

3. Point adjustment: definition of the origin for x, y, and z. The origin of the z-axis through the measured x–y plane and the origin for the x- and y-axes through the intersection of lines A and B are described in Fig. 8.27.

The probe element must be chosen so that all of the elements to be measured are reachable. For this reason, more probing elements may be necessary, as shown in Fig. 8.28. One calls the connection of more than one probe elements a probe combination or a probe tree. It can also occur that different sensors are needed for a particular measuring task. In the preparation for the calibration, a probe element is specified as a reference probe element. With multiple probe combinations or sensors, there is only one reference probe. Whenever more probe combinations are necessary, the work is implified by an automatic probe-changing mechanism.

The workpiece is set or clamped onto the measuring machine table, so that it lies as firmly and clearly as possible. Easy clamping elements, clamping element component systems, or special clamping devices are available. All of the form elements to be measured must be reachable with the chosen probe configuration. It is not allowed to come into contact with gripping or clamping elements.

The specification of the sequence of probing occurs essentially from an economical point of view (shortest routes between the elements). The type of probing (for example, mechanical or optical, point-to-point or scanning, auto-centering) is substantially limited by the technical capabilities of the measuring machine. Remark: Scanning is the independent tracing of a workpiece contour by the measuring machine. One must give a starting point, end point, the direction, and a scanning plane. The path results from the cut of the scanning plane with the workpiece surface. The scanning speed and point density must be provided by the operator. The minimum measuring point court follows from the type of geometrical elements (see Sect. *Basics*). The aim of measurement (parameter or form deviation) forms the basis for the definition of the real used measuring point count. With increasing numbers of measuring points, the determination of the individual form elements becomes more certain. In any case, more measuring points than the minimum should be utilized. Around 2 to 3 times the minimum point count for the recording of a measurement and its deviation is sound from a measuring technology and economic viewpoint. For form deviation, one needs substantially more points. Here, the number depends on the size of the smallest geometric portions to be recorded. The measuring points should, if possible, be evenly distributed on the measured element and in addition must be able to represent these form elements with the desired deviations.

Phase 2: Preparation and Measuring. The probe combination is assembled as defined and then calibrated. Figure 8.28 shows the calibration of one probe of a two probe combination. Through measurement of a calibration standard (a sphere with a known diameter and a very small spherical deviation) with both probes, the actual diameter of the probe tips and the distance of the middle of the sequential probe sphere to the middle of the reference probe sphere in the x, y, and z directions determined. After clamping the workpiece the machine parameters (e.g., velocity, measuring force) are to define.

The degree of automation of the measuring machine is of deciding importance for the probing of the measuring points. For manually controlled machines, the entire course (measuring point probing and manoeuvres in space between the machine elements and gripping devices) at every single point is realized by the operator. CNC-controlled machines allow the programming

of a course of motion and have an automatic drive to adopt the required positions. One distinguishes between three possible programming methods:

- By teach-in, the machine learns the course of the measuring directly from the handling of the operator, i. e., his motion. He executes the course to the points to be measured as on a manually controlled machine. In this way all of the measuring point coordinates and all necessary points for collision-free manoeuvring between the machine parts and gripping mechanisms can be recorded. Afterwards, the control system can independently reproduce this course.
- The generation of the target coordinates can take place far from the machine. In this method, all measuring coordinates and position coordinates are derived from the figures or drawing data. This procedure requires great spatial imagination from the programmer. This approach is used mostly in conjunction with teach-in portions.
- Computer-aided design (CAD) systems that have a measuring module at their disposal can directly provide measuring machine programs (compare with next section).

Phase 3: Evaluation. The description parameters for the associated geometric elements are determined from the measuring points through compensating calculations (for example: Gauss, inscribed circle, circumscribed circle). Take notice that either the measuring points or the description parameters are to correct by the calibration data (probe sphere diameter, distance to the reference probe). On the foundations of the description parameters, elements can be mathematically linked or joined with each other. So new characteristics that describe the geometry of the entire workpiece can be developed. The parameters of the individual elements (diameter, length) and the resulting parameters from the links (distance, angle) can be compared with the nominal and tolerance data. Afterwards, the results can be represented graphically or numerically.

Integration into CAD and CAM

Coordinate measuring machines (CMM) are suitable for integration into CAD/CAM environments. This integration is possible under two criteria.

In CAD systems, the geometric data for a new workpiece is created. With that, outstanding conditions exist for the derivation of the measuring program on the basis of this data. In addition to the data describing the workpiece, a measuring module implemented in the CAD system needs data for the possible elements of the probe combination and the gripping system. *The complete measuring program can be produced in a simulated measuring run.* It is important for later functions that the coordinate system for the description in the CAD system agrees with that for measured workpiece on the measuring machine. The measuring program can be written in a specialized programming language for the measuring equipment manufacturer or in the manufacturer-independent universal language dimensional measuring interface specification (DMIS), depending on the available measuring modules. For integration into the measuring machine program a special interpreter is then necessary. After successful measuring, the possibility exists for the acquired deviations to be transferred back to the CAD system, to be processed or represented. Of course, an external process through best-fit and distance determination of the CAD data is possible.

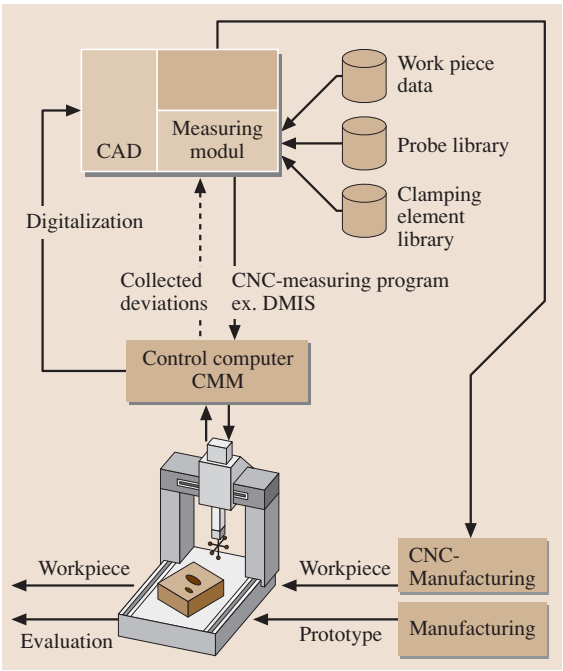


Fig. 8.29 Connection CMM–CAD

The second possibility for the integration of a coordinate measuring machine is the *digitalization of an unknown workpiece geometry*. Here a manufactured sample forms the basis (prototype). This sample is touched in a previously defined grid. The measuring

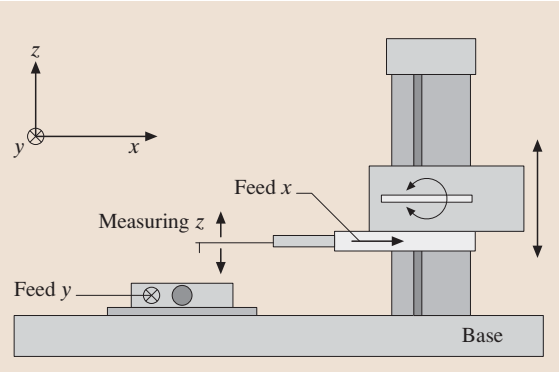


Fig. 8.30 Stationary surface metrology instrument

point coordinates acquired are handed over to the CAD system. There, a surface feedback or reproduction of the data model takes place. This can then be transferred to a manufacturing machine that makes a copy of the prototype.

8.2.7 Surface Metrology

Instrument Technology

The surfaces of geometric objects continuously provide an informative testing object. Considerations include appearance, the evaluation of the expected functional behavior, and manufacturing and wear conditions. The application of modern measuring devices in conjunction with sophisticated computer technology facilitates detailed evaluations and conclusions, which only a few years ago were unthinkable. With the available equipment and the demands of the user, these approaches are being widely applied.

Small, independent, portable instruments with skid pick-up systems are easily implemented to directly monitor tooling processes. In addition, stationary laboratory instruments with free probing systems, with correspondingly large measuring ranges and higher resolutions, have become established. Figure 8.30 shows the basic construction of a stationary surface metrology instrument. It consists of a base and pillar. The feed unit, attached to the pillar, can be positioned aloft and be pivoted opposite the base. The probing system is positioned with high precision on the feed unit. This positioning constitutes the measuring basis for the probing of a measuring object. If required, the system can be supplemented with an additional y-table. This y-table facilitates the step-by-step movement of the workpiece perpendicular to the feeder movement enabling three-dimensional scanning.

Instruments constructed as component systems allow a surface probing system to be used as a contour probing system (larger measuring range, many different types of possible probes) and therefore, to expand the application range of the instrument considerably. It is also possible to implement, instead of a mechanical probing system, a noncontact probing system if the surfaces are sensitive to mechanical probing. One option is to use auto-focusing laser sensors (see capitel laser measuring technology).

Developments in electronics over the last few years have made it possible to digitize large quantities of data with a higher decimal precision (higher number of levels) with sufficient speed. This has enabled the collection of many measuring points (with a smaller distance between measuring points at an acceptable collection speed) over larger measuring ranges with greater resolution (small Δz). The accessories for roughness measuring instruments range from stylus stop attachments to a large selection of speciality probing elements and various skids. The instrument can be tailored to the measuring task with these specialized accessories. As well as instrument equipment, the environmental conditions are also an important aspect to consider. The implementation in the laboratory setting, in contrast to the factory floor, brings with it a significant improvement in environmental conditions. This facilitates tasks with greater demands for precision.

ISO Standards and Consequent Requirements

The relevant ISO-standards give not only the definition of characteristics but also requirements on measuring instruments (e.g. probe tipradius; distance between measuring points) and software (e.g. phase-correction Gauss filter). The profile, which is obtained by means

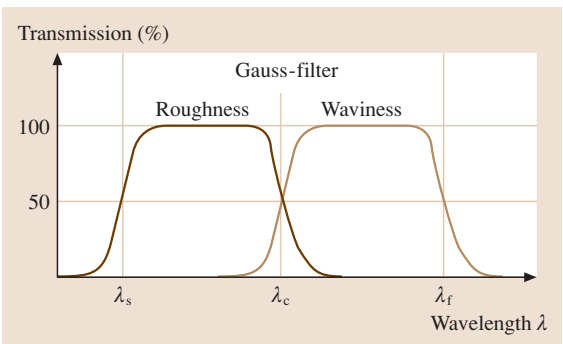
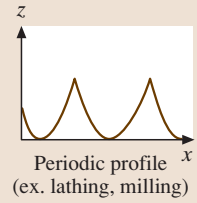
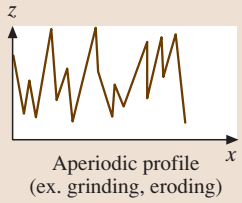


Fig. 8.31 Roughness and waviness profiles



Periodic profile
(ex. lathing, milling)



Aperiodic profile
(ex. grinding, eroding)

Average ridge widh of the roughness profile *RSm* (mm)	Measuring conditions: Threshold wavelength λ_C (mm)	Threshold wavelength λ_S (µm)	≈ Ratio λ_C/λ_S	Single measured length *lr* (mm)	Measured length *ln* (mm)	Probe length *lt* (mm)	Maximum tip radius r_{tip} (µm)	Maximum point distance Δx (µm)	Arithmetic mean of roughness *Ra* (µm)	Maximum profile heigh *Rz* (µm) *Rz1*$_{max}$ (µm)*
0.013 < *RSm* ≤ 0.04	0.08	2.5	30	0.08	0.40	0.48	2	0.5	(0.006) < *Ra* ≤ 0.02	(0.025) < *Rz* ≤ 0.1
0.04 < *RSm* ≤ 0.13	0.25	2.5	100	0.25	1.25	1.50	2	0.5	0.02 < *Ra* ≤ 0.1	0.1 < *Rz* ≤ 0.5
0.13 < *RSm* ≤ 0.4	0.80	2.5	300	0.80	4.00	4.80	2; 5	0.5	0.1 < *Ra* ≤ 2	0.5 < *Rz* ≤ 10
0.4 < *RSm* ≤ 1.3	2.50	8	300	2.50	12.5	15.0	5	1.5	2 < *Ra* ≤ 10	10 < *Rz* ≤ 50
1.3 < *RSm* ≤ 4	8.00	25	300	8.00	40.0	48.0	10	5.0	10 < *Ra* ≤ 80	50 < *Rz* ≤ 200
Additional for *RSm* to all profiles	Phase correct Gauss-filter						∢ 60° (90°) 0.75 mN		For *Ra*, *Rq*, *Rsk*, *Rku*, *RΔq*	For *Rz*, *Rv*, *Rp*, *Rc*, *Rt* (*For *Rz1*$_{max}$ …)

Fig. 8.32 Measuring conditions (ISO 3274 and ISO 4288)

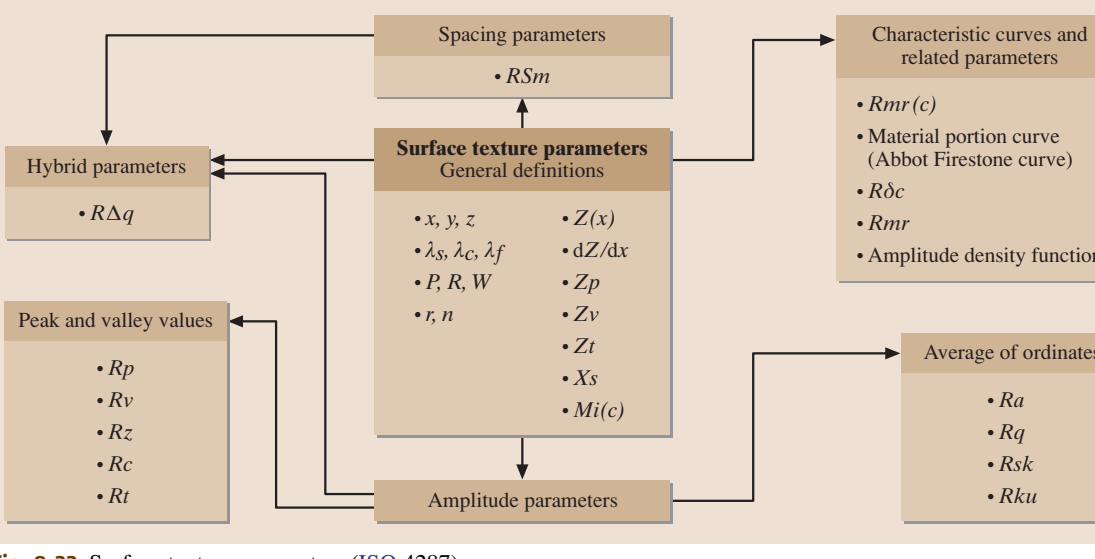


Fig. 8.33 Surface texture parameters (ISO 4287)

of the section probing method, is called, after the application of the filter for short wavelengths λ_s, the primary profile (*P*-profile). The roughness profile (*R*-profile) is obtained through the deletion of the long-wavelength profile features (threshold wavelength λ_c) from the primary profile. The waviness profile (*W*-profile) is made by filtering the primary profile by means of λ_c and λ_f, as depicted in Fig. 8.31.

The threshold wavelengths λ_c and λ_s necessary for this filtering readable in Fig. 8.32 after profile classification between periodic and aperiodic. No concrete definition currently exists for the threshold wavelength

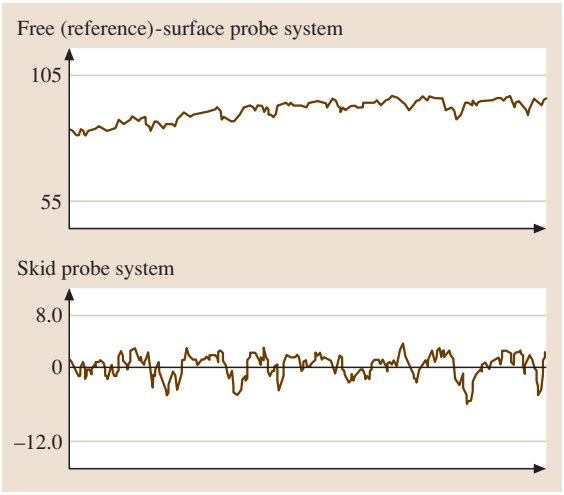


Fig. 8.34 Primary profile

λ_f, only the recommendation of $\lambda_f = 10(5)\lambda_c$. Besides the threshold wavelengths for the separation of profile elements, definitions of the maximum probe element radii and the distances between measuring points are being established. However, there are severe restrictions for the use of instruments in this way.

Characteristics can be calculated from all profile types (P, R, W). Figure 8.33 shows the arrangement of surface characteristics based on the example of the roughness profile. Fundamentally, horizontal, vertical, and hybrid characteristics can be differentiated. These are supplemented by characteristic curves from which parameters can be derived.

Because the explanation of the single characteristic definitions would be too lengthy, the corresponding standards (ISO 4287) is referred to instead.

Analysis of a Surface

Whereas the previously explained, easy-to-handle instruments with skid pick-up systems can only collect and evaluate roughness, instruments with reference-surface probing systems produce data that, with proper software, can measure not only roughness and waviness characteristics but also size, form, and positional deviations. Also, a description of surface alterations by abrasion and coating, amongst other effects, is possible with these systems.

The difference in the information content of the data is immediately recognizable in Fig. 8.34. Whereas the skid pick-up system only approximately captures the roughness portion, a slant and curvature can be recognized in the data collected from the reference-surface

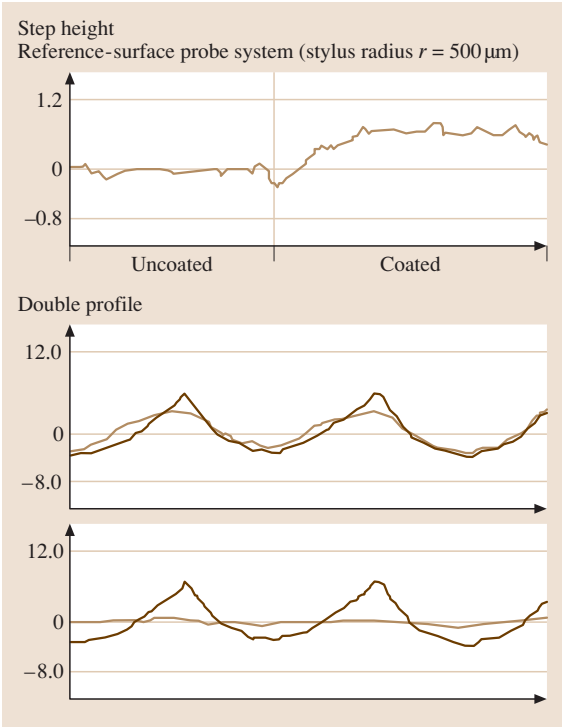


Fig. 8.35 Application for the manufacturing measurement technology

probe system. The skid pick-up system manages with essentially small measuring ranges and poor guidance, which makes these instruments much more affordable. However, one should pay attention to the fact that the mechanical filtering of the signal, by means of the skid, cannot be removed during the follow-up evaluation.

The larger information content collected with a reference-surface probe system is accompanied by substantial disadvantages related to the instrument technology. The disadvantages are the relatively large measuring range of the probing system, which is required in order to obtain the primary signal within a justifiable adjustment effort and the fact that the straightness of the feed unit movement must be very good, as it contributes to the measuring signal. Both effects lead to relatively high instrument costs.

Figure 8.35 shows the example of the implementation of surface metrology equipment with a reference-surface probing system. On the top, data collected from a partially coated workpiece is shown. The implementation of a probe element with a comparatively large radius ($r = 500\,\mu m$) leads to the low-pass mechanical filtering. For the compensating calculation, only data

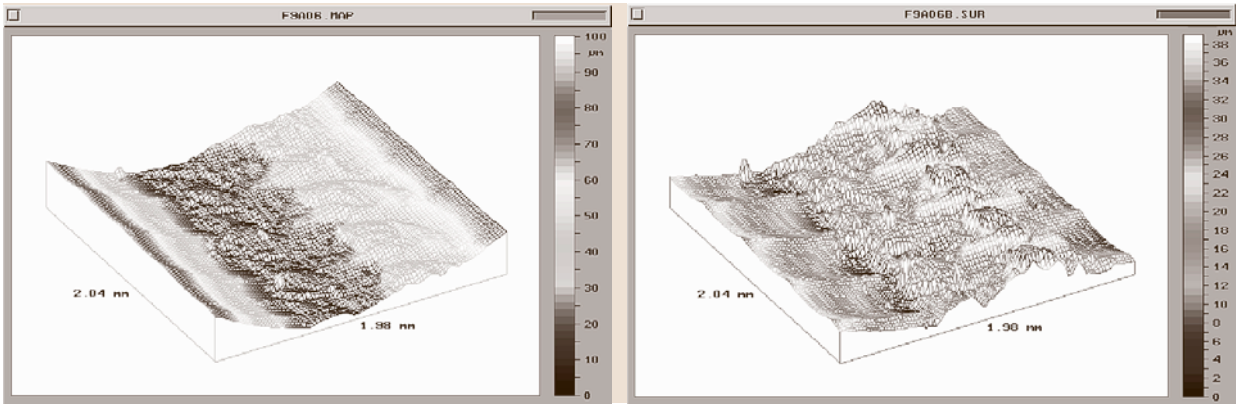


Fig. 8.36 3-D surface structure

that represents the uncoated material is used. After this calculation, the course of the layer thickness between the measured profile and the extended compensation profile is clearly recognizable in the coated section. Similar approaches are also suitable for wear measurement whenever unworn workpiece sections are still available. On the lower part of Fig. 8.35, production-progress is shown by a double profile. The basic result is a turning surface profile, which is then smoothed in a second processing step. Clearly, the varying manufacturing results achieved by changing the machining parameters (low or optimal pressing force on the tool) can be noted.

The additional assembly of *y*-shift table perpendicular to the actual feed direction (Fig. 8.30) allows the collection of data from flat, three-dimensional structures. By the use of an appropriate software package, one can derive three-dimensional surface characteristics from this data, or rather a visual impression of the surface for the benefit of the user. This allows conclusions on properties of the surface, which one cannot easily derive from a single profile. Figure 8.36 shows the three-dimensional measured structure of a workpiece that was milled with a spherical-headed milling tool. In the left-hand side of the figure, a detail from the actual surface, is shown. The differences in altitude are represented by different brightness. There is also the choice of an isoline representation (connected lines of equal height). The curvature left behind by the spherical cutter is clearly recognizable. Under this relatively strong curvature, the detailed structure remains hidden.

The right-hand side of Fig. 8.36 shows the same section. In order to make the details clearer, the dominating curvature in the left part of Fig. 8.36 is removed with the use of a compensating calculation (a three-dimensional second-order polynomial fit), Now the milling feed is clearly recognizable. The unclear structure in the middle of the illustration emerges because there are no definite cutting properties in the domain of the center of the cutting tool.

8.2.8 Form and Position Measuring

Instrument Technology

Form inspection equipment in most cases consists of a base and a column. A turntable is integrated into the base. On the column, an arm is located, which carries the measuring system. Even after construction of the instrument, the turntable, the column, and/or the arm can be arranged as a measuring base. The measuring system is in most cases inductive. The probing is carried out with a spherical probe element. In order to keep abrasion to a minimum, ruby probing elements, as used in

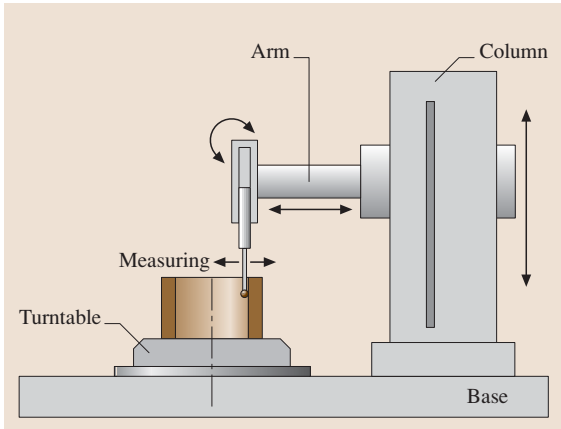


Fig. 8.37 A form inspection instrument

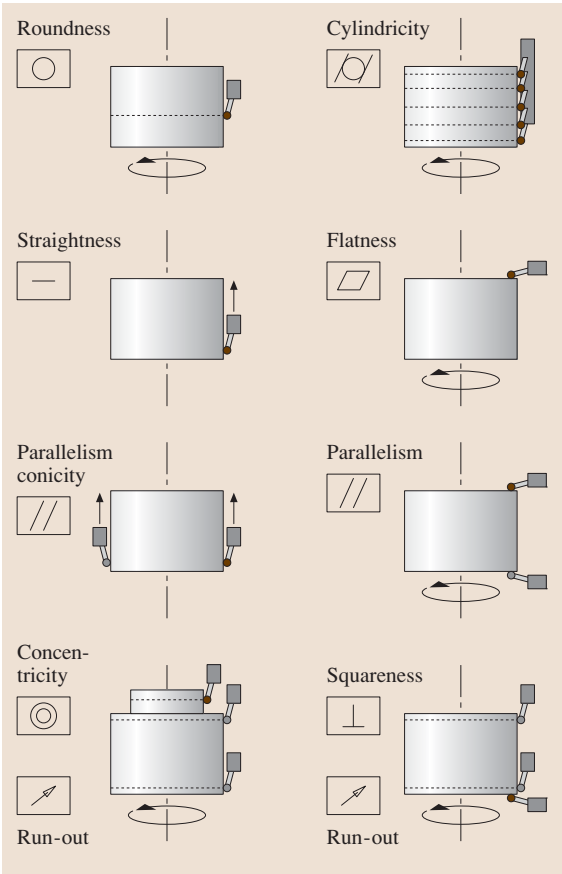


Fig. 8.38 Measuring execution

coordinate measuring machines, are utilized. The measuring system is inclinable so that the direction of the measuring force can be reversed. Therefore, almost all necessary measuring points can be reached. Figure 8.37 shows a vertically arranged probe system, by which the measuring force from the workpiece acts radially outwards. Subject to the measuring task, one or more feed axes can be implemented.

Measuring Execution

Figure 8.38 shows, in the upper left-hand corner, a roundness inspection. After positioning the piece on the turntable, the piece is orientated to the rotating axis so that the measuring system remains in the measuring range during the course of inspection. By a large measuring range, visual judgement suffices for orientation. The large measuring range leads to a bad resolution. A more-accurate orientation of the piece is demanded for a smaller measuring range, leading directly to a better resolution.

After the probe is brought into contact with the desired position on the object, the measuring object is rotated 360° by the turntable and, while the object is turning, the measuring points are scanned. An ideal circle is calculated from the measuring points. For this, depending on the aims of the inspection, compensating methods (for example, Gauss, circumscribed circle, inscribed circle, see Sect. 8.2.6) can be chosen. In reference to the compensating circle, the range between the lowest point and the highest point is the roundness deviation. For the investigation of causes of deviations, one can filter the desired frequency domains or use a frequency analysis e.g. a fast Fourier transform (FFT) (Fig. 8.39).

In the left-hand side of the figure, the unfiltered circle form deviation is depicted for evaluation of the

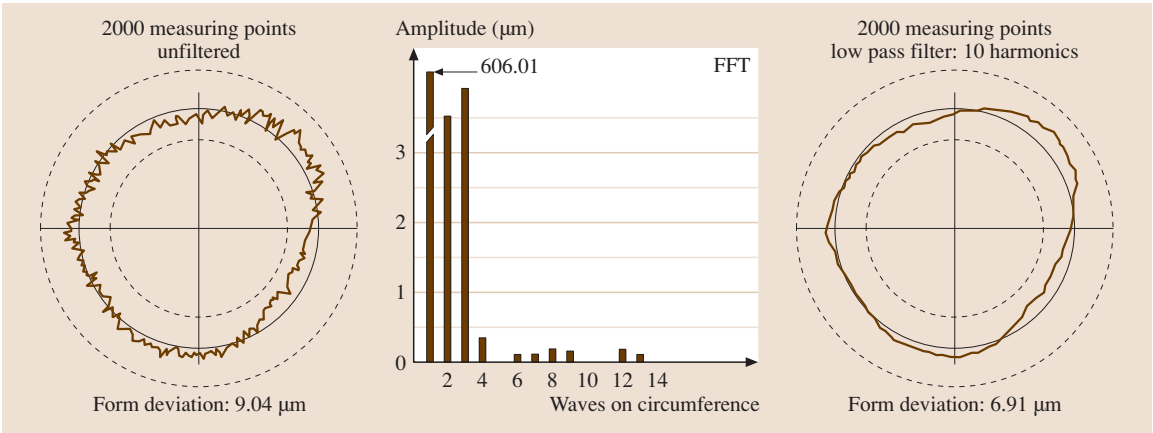


Fig. 8.39 Roundness deviation

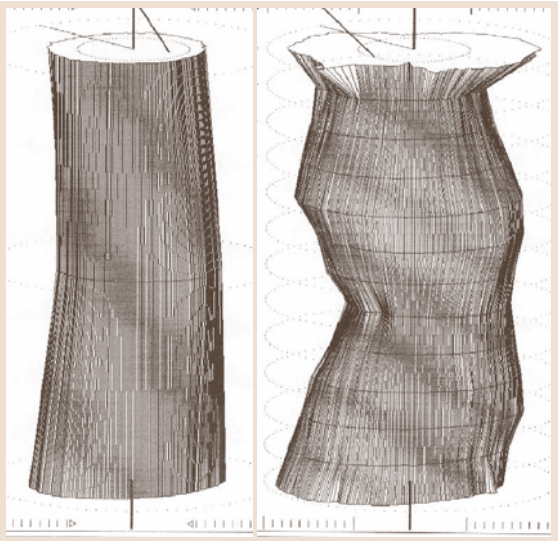


Fig. 8.40 Cylinder form deviation

workpiece. The frequency analysis in the graph next to it shows that waves, especially with a period of 2 and 3 harmonics, dominate the result. The right-hand side of the figure shows the measuring result after running the data through a low-pass filter (limit wavelength of 10 harmonics). Here, it is easier to recognize causes of deviations and to implement corresponding quality-orientated corrective action into the manufacturing process.

Figure 8.40 shows the collection of cylinder form deviation. Here, in addition to a high-precision rotating axis, a column axis with correspondingly small straightness deviation is used in order to facilitate vertical movement. When positioning, movement of the arm should be avoided. All of the collected measuring points, located on circles, are put through a compensating calculation together. From the distances between the single measuring points and the compensating cylinder, the cylinder form deviation is derived.

Just as the number of points on a circle is important for the determination of the circle form deviation, so the number of measured circles for the cylinder form deviation is not to be neglected. Figure 8.40 shows, on the left, a picture that used three circles and on the right, a picture that used 13 circles to calculate the cylinder form deviation. Clearly, there is a noticeable difference in information content between the two.

To detect referenced elements, like for example, the concentricity depicted in Fig. 8.38, one uses the methods already known from coordinate measuring technology. First, both of the circles on the lower cylinder are probed. The circles should have the largest possible distance from each other in order to reduce the measuring uncertainty. Compensating circles are calculated from the corresponding measuring points. The line that connects both mid points of the circles serves as the reference axis. After the inspection of the circle on the smaller cylinder and the corresponding determination of the parameters of the compensating circle, the eccentricity is calculated as the distance between the midpoint of the circle and the reference axis. The tolerance zone for the concentricity describes a circle around the reference axis.

8.2.9 Laser Measuring Technology

Basics

Based on its properties, laser measuring technology has gained a strong position in the field of integrated measuring technology inside the manufacturing processes. In this section, a few examples are explained and the advantages and disadvantages highlighted. The application profiles of laser measuring systems are as different as their parameters, and range from simple dimensional completeness checks to high-precision measuring systems that determine surface roughness or even the acceleration of tooling machine components. From the many possible implementation variations, only a few are considered, those that use the various properties of light. Next to the time-of-flight method, which is not commonly implemented in mechanical engineering applications, come uses of such properties include linear propagation, reflection, or interference. The selection of the best sensor for the specific application and the arrangement of the sensor to the measuring object is based on criteria such as:

- Operating distance
- Measuring range
- Resolution (scale value)
- Measurement uncertainty
- Measuring spot diameter
- Light color and intensity

However, the price of the entire system, which consists of sensors, processing electronics and possibly components for the movement of the measuring object or the lose, is also quite relevant.

In addition to the technical and scientific considerations, protection against accidents is also an issue that should not be ignored. Because most laser meas-

uring systems are equipped with low-power lasers (hazard category 1–3), skin burns are rather improbable. However, the eyes must be well protected. Reflected radiation from metallic surfaces is also problematic. The mandatory guidelines, for instance in Germany, demand labeling for hazard category 3B and above, the application of opaque enclosures, laser protection eyewear, and special warning and emergency facilities.

Laser Applications in Measuring Technology

The *linear propagation* of a laser beam is used by a whole range of measuring devices. To clarify this application, two examples are presented. The first is a laser scanner (Fig. 8.41), which amongst other applications is used here for the quick scanning of the diameter of shafts.

The laser beam, created by a diode, is expanded into a fan shape and aligned so that the beam is parallel. As shown in the figure, this alignment is done by optical components. There are also laser scanners in which the beam expansion occurs by means of a spinning polygonal mirror. The laser beam is directed toward the measuring object. Behind the measuring object, a receiver is arranged where the laser beam strikes a light-sensitive CCD line array. Because the laser beam is broken by the measuring object, in this case the shaft, the size of the shadow on the receiver can determine

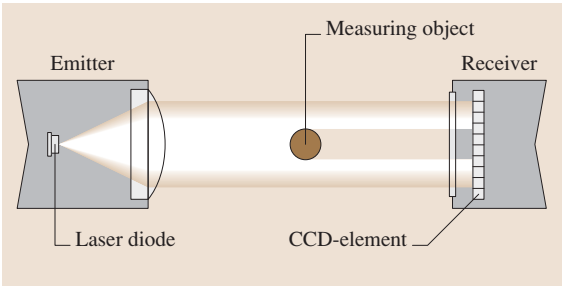


Fig. 8.41 Laser scanner

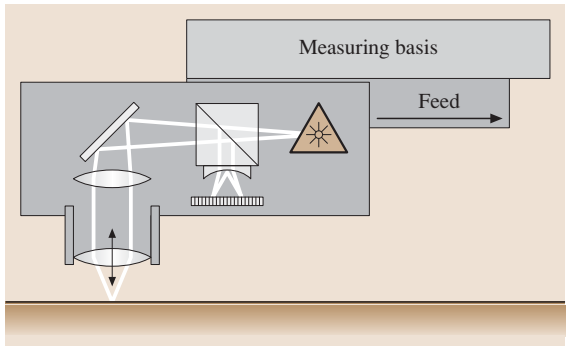


Fig. 8.43 Autofocus sensor

the diameter of the object. The emitter and receiver can be adjusted axially to the shaft very quickly, so that in a short time, the diameters of various shaft sections can be determined.

A second application of linear laser expansion is shown in Fig. 8.42. The emitter, which emits a laser beam with a narrow diameter, is arranged on the fixed part of a tooling machine base. The receiver is assembled on the sliding bed and consists of a CCD matrix sensor field (in the middle of the figure). Different manufacturers also use four quadrant diodes at this location as a receiver.

Now, the sliding bed of the tooling machine is steered to its home position. The point at which the laser beam strikes the sensor field is registered (y_1, z_1 at $x_1 = 0$ mm). On the way to the end position, it is stopped at defined points and for each stop the laser point is registered on the sensor field (y_2, z_2 at x_2, and so on). From the obtained points, the straightness deviation of the sliding bed movement is shown divided into the x–y and into the x–z planes.

A laser beam reflected by the measuring object is generally used by laser distance sensors. These are employed for the contactless measurement of deformable, very rough, sensitive, hot, or moving surfaces. A laser

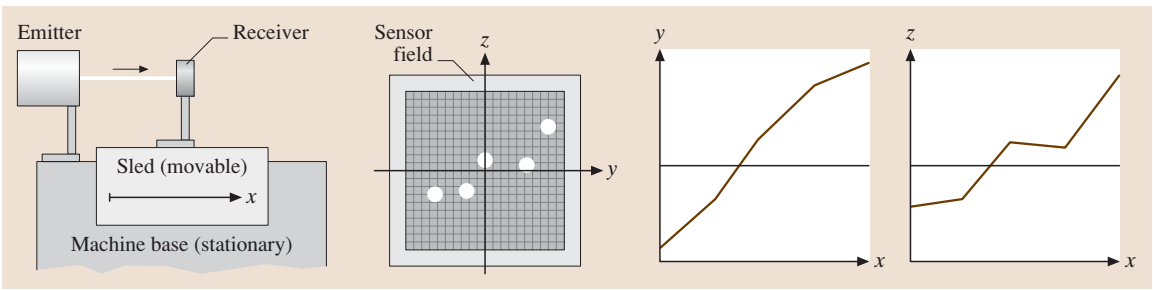


Fig. 8.42 Alignment measuring system

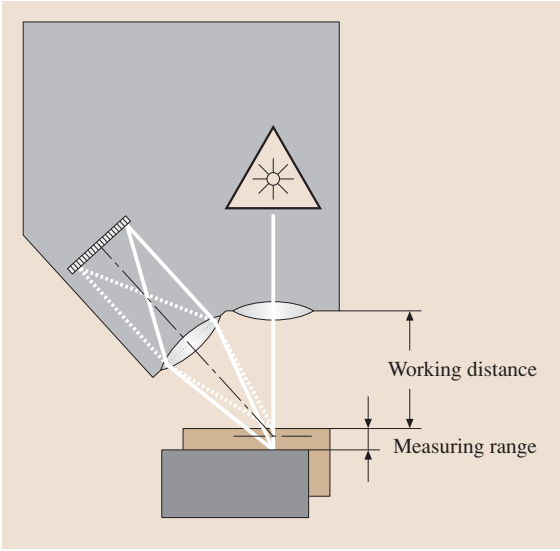


Fig. 8.44 Triangulation sensor

autofocusing sensor and a laser triangulation sensor are given as application examples. Figure 8.43 shows schematically the construction of an autofocus sensor (Mikrofokus from UBM). The laser beam is projected onto the surface of the measuring object and, from this surface, the laser beam is reflected. Specialized electronics monitor the focusing and, when necessary, readjust the lens system. The displacement of the lenses is recorded as the measuring value. The previously mentioned sensor example has a working distance of about 5 mm and a measuring range of 1 mm. The focused laser spot on the surface of the measuring object has a diameter of 1–5 µm, while the beam has an angular aperture of 30°. The electronics used in the example, allow the measuring signal to be broken up into steps of 16 nm. With these parameters, the sensor is well suited for contactless surface inspection. The advantages of noncontact probing are also accompanied with disadvantages. Problems especially arise due to locally strong reflective (for example, single grains on grinding wheels) and also by nonreflective materials as well as by steeply inclined surfaces (over 30°). In these cases, the reflected beam cannot be collected for evaluation. The reaction of the system to such problems is variable (measuring value setting, holding or searching) and must be take into consideration by evaluation of the signal.

Triangulation sensors (Fig. 8.44) likewise project a light point onto the surface of the measured object. This light point is observed from a defined angle. By changing the distance between the sensor and measuring object, the reflected point wanders across a CCD row. In this way, the change of the distance is collected directly. In addition there are some sensors project one or many laser lines onto the inspected surface. These lines can then be evaluated in terms of the contour of the measuring object by means of a CCD matrix.

The chosen example sensor has an average working distance of 105 mm, a measuring range of ±25 mm, a projected laser spot diameter of 0.1 mm, and a triangulation angle of 18°. With these parameters, these sensors can be applied, for example, to:

- Size or form inspection
- The collection of runout and position deviation
- The collection of thickness
- The collection of deformations
- Completeness or integrity control

As Fig. 8.45 shows, single or combined sensors are necessary for these applications. The measuring signal collects the surface of the measuring object in the range of the entire measuring spot diameter. So a low-pass filter processes and determines the mean distance in this

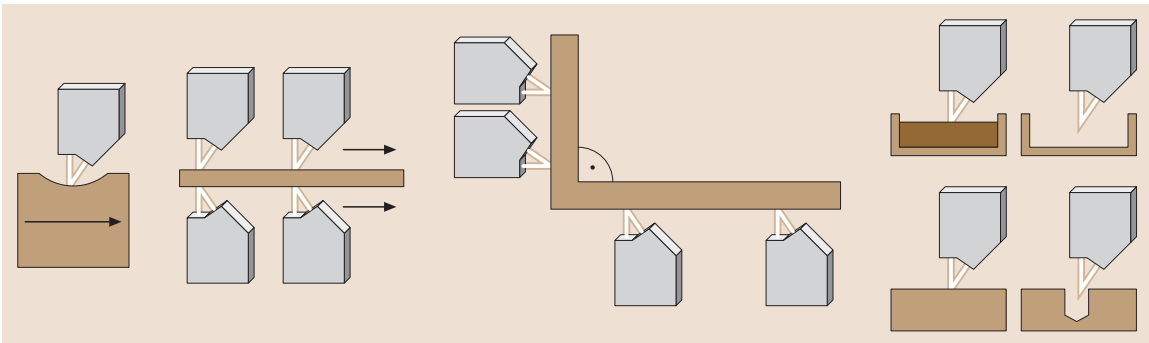


Fig. 8.45 Examples of applications using triangulation sensors

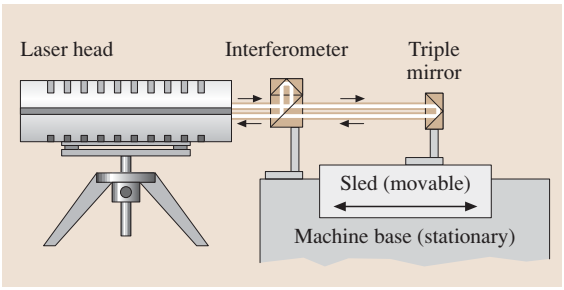


Fig. 8.46 Interferometer for the determination of position

area. The plane in which the beam is sent and reflected is arranged parallel to the surface structure to guarantee a reflection.

In laser interferometers, the laser beam is separated into two beams (reference and measuring beam) by means of a semireflecting mirror. While the reference beam always follows the same path inside the interferometer, the measuring beam is reflected off a triple mirror on the object to be measured, as shown in Fig. 8.46. When the two beams come together, interference occurs (Fig. 8.47). When the peaks of the waves meet (phase shift 0°), the light is amplified. When the peak of one wave meets the valley of the other (phase shift 180°), the waves cancel. The distance between the interferometer and the tripel mirror changes when the sled moves. The phase between the reference and measuring beam shifts. This resulting change in the interference is interpreted as the measuring value.

This type of measuring system is used for applications with large lengths (10–20 m) and with a high resolution (5 nm). It only collects, in each case, the changes from an initial state. A disadvantage is that the optical components must be mounted on the object that is to be measured and that the frequency of the light, and therefore the measuring result, is strongly influenced by environmental conditions such as air temperature, humidity, and pressure. Corrections can be made by collecting these environmental conditions and carrying out a subsequent adjustment of the data, or more specifically, through the application of a refractometer (an instrument in which the effects of the environmental

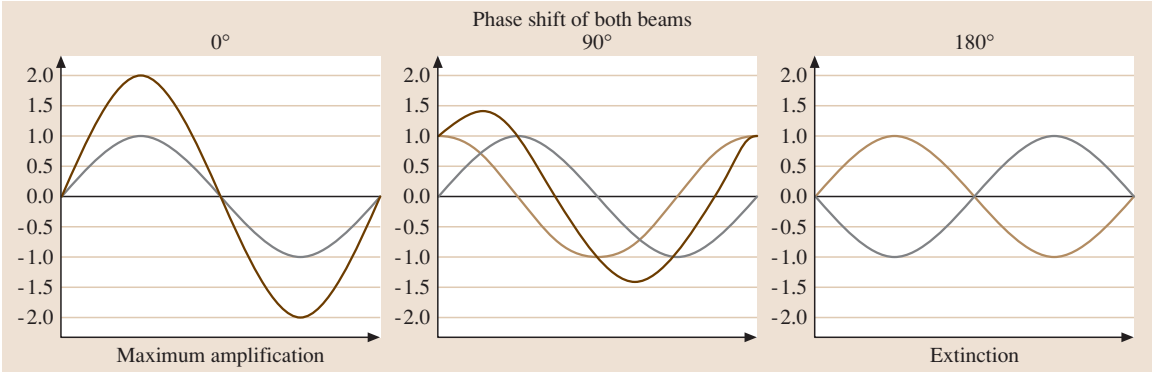


Fig. 8.47 Phase shift and interference

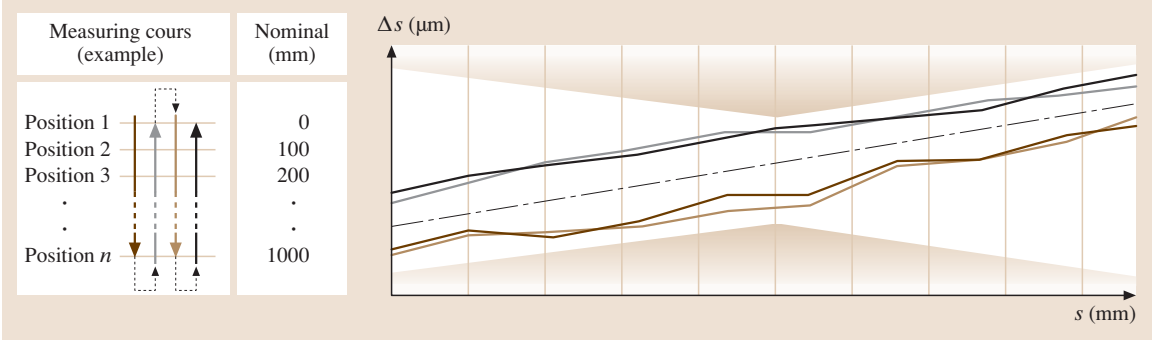


Fig. 8.48 Position deviation, measuring course and results

conditions can be directly compared with the light properties in a vacuum).

Figure 8.46 shows an example combination that determines the position of a machine component in respect to another, or rather to a chosen fixed point. This set-up is used as an independent measuring system (for example, with coordinate measuring machines) but also as a measuring system for the collection of position deviations. Figure 8.48 shows the procedure for the collection of position deviations of a tooling machine. While the machine, controlled by its measuring system, follows the defined positions labeled 1 to n, the position deviations are determined by means of a laser interferometer. The measuring course, which is step-by-step executed over the entire range of to be measured positions, is completed multiply in both directions (increasing and decreasing values). The deviances collected in this way are shown on the right-hand side of Fig. 8.48. From the depicted course of measuring values, a linear trend and span superimposed with random sections can be read. The limits, shown in the diagram, depict usual tolerance ranges for tooling machine position deviations. The systematic parts are compensated for, when necessary, through mechanical aligning of machine components, or afterwards by means of a correction table or correction function.

Similar arrangements allow the collection of tilt and rotation angles as well as the determination of straightness deviations. When the measuring value transmission frequency can be chosen to be sufficiently large, a dynamic series (for example the reaching of a desired position) can be collected with the layout shown.

8.3 Measuring Uncertainty and Traceability

The result of measuring is the measurement value. This measurement includes the true value and also systematic and random deviations of the measurement. It is possible to estimate the expected value of the measurement by calculating the arithmetic mean of several independent measurements. The more single values that are processed, the more the random parts of the measuring deviation are reduced. The systematic part of the measuring deviation can be determined by measuring an object with a known value. Such objects, measurement standards, are measuring blocks for lengths. Angles can be represented with angle gauges, precision polygons, or with sine bars used in conjunction with measuring

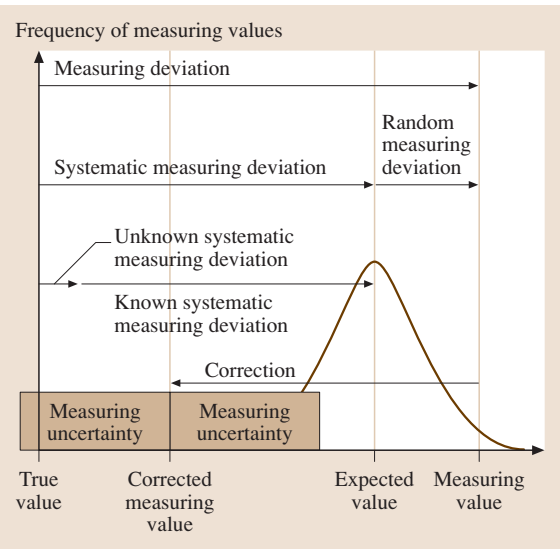


Fig. 8.49 Summary of measuring values

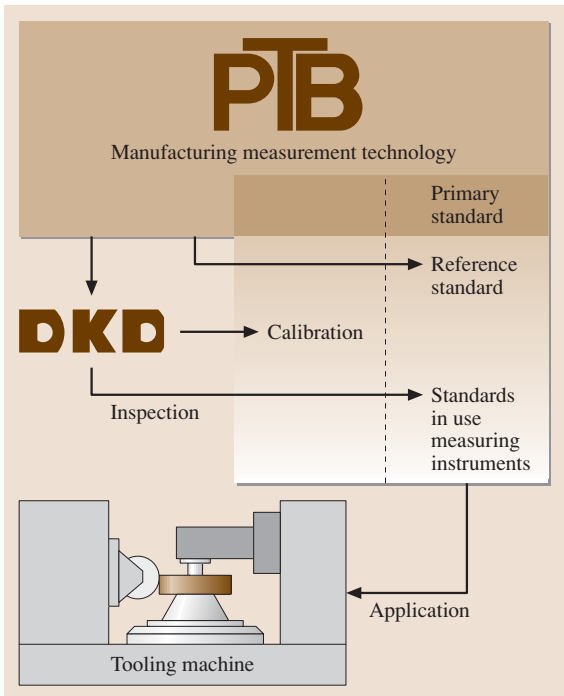


Fig. 8.50 Traceability

blocks. The known value of the measurement standard (whose small deviation from the true value is negligible for the purpose of comparing), is called the correct value. The systematic measurement deviations can be calculated from the difference between the measured value of the measurement standard and its correct value. Afterwards one can correct all measuring values by use of the known systematic deviation. The non-measurable or not measured (for instance too expensive) elements of the systematic measuring deviation are combined with random measuring deviations to constitute the measuring uncertainty. So, the measuring uncertainty is a parameter obtained from the measurement. It describes the region around the corrected measuring value where the true value must be found. The complete result of measuring is given as the corrected measuring value plus/minus the measuring uncertainty.

In order to guarantee the correctness of measuring results, measuring devices must be affiliated to the national standard of the respective measurement. In Germany for instance, this is the National Physical Technical Institute (PTB), located in Braunschweig, which are responsible for the representation and propagation of physical units. Through the PTB, calibrating laboratories are accredited, so that their measuring devices and measuring standards coincide with the national standards (the units), according to the defined and accepted techniques. These are the laboratories of the German Calibration Service (DKD).

All in-company measuring devices, or rather standards, should consistently be traceable back to the national standard of the PTB. In order to verify this, the instruments are calibrated by a DKD laboratory or the PTB themselves. *Calibration* is defined as the inspection of measuring devices and measurement standards with reference to the accepted national standard. The successful calibration is generally documented through a protocol, the calibration certificate. On the calibration certificate, all of the calibration results, the reference standards, and additional measuring equipment (used during the calibration), the environmental conditions, and the calculated measuring uncertainty are documented.

One should use certified instruments, standards, and methods in order to achieve cost-effective, in-company control of inspection instruments. Otherwise, a traceability certificate is not possible.

The setting-up and balancing of a measuring instrument, by which known systematic deviations of the measuring result are eliminated, is called *adjusting*.

8.4 Inspection Planning

Inspection planning means the planning of quality inspection in the entire production process, from the arrival of raw products until the delivery of the final product. For this, inspection tasks and procedures are specified with inspection feature, inspection location (close to production, measuring room), frequency of inspection, point of time within the production process, inspection methods, inspection equipment, and operators. One should consider both technical and economic aspects. The inspection planner must consider knowledge of the function and application of the piece or the components, safety hazards, the production process, technical documentation (drawings, standards, stipulations), and the inspection equipment. It should be consistently checked that the data is complete, current, and inspected (approved by the operating department).

For the choice of inspection procedures, a systematic search in the drawing(s) ordered by

- The type of parameter (measure, form and position tolerance, surface tolerance, etc.)
- The drawing view
- Grid squares (for example, upper left starting clockwise)

in the work plan according to maintenance sequences, in the documentation, delivery instructions, and contractual arrangements, is essential. The definition of the inspection frequency (number of samples, size of sample) occurs on the basis of mathematical statistical facts. The time point in the production process allows company organizational and economical considerations. The late recognition of inadmissible errors can bring about several disadvantages.

The definition of inspection methods and inspection equipment are related and thus should each be chosen with the other in mind. The choice of measuring device begins with consideration of the required information content of the measuring result. In this way the aim of the inspection (evaluation of workpiece or process, manufacturing control) and the impact of the measuring instrument itself can be taken into

account. Geometrical limitations, such as the accessibility of the piece, the geometry of the probing element, the range of measurement (direct measuring, difference measuring) of the instrument, and especially for soft materials, the measuring force, are previously decided. Finalized statements for the usability of a measuring instrument can be made after inspection of the scale division value and the measuring uncertainty. The measuring uncertainty must adhere to the ratio to the inspected tolerance by the relation $\frac{u}{T} = 0.1$ to 0.2. Alternatively, it is possible to obtain measuring capability coefficients and check the adherence of these characteristics to previously defined limits. Supplemental criteria are, for example, the surfaces available for the measurement, and transfer for processing, protocol and archiving.

The required measuring time (capability of the measurement to be automated) in conjunction with the number of pieces to be tested is the essential criteria for the cost effectiveness of the application of a measuring device. Included in the inspection costs are also the equipment costs, equipment observation, calibration, and personnel costs (work time, education).

To guarantee the comparability of the measuring result and low uncertainty of the acquired characteristics the following conditions should be taken into account when specifying measurement methods:

- Explicit guidelines for the measuring procedure including, the required parameters for an appropriate measurement. Specification of the reference basis for the measuring procedure, to the accuracy of the applied measuring instruments and measurement standards, gripping elements employed, and additional measuring equipment.
- Details of the measuring strategy, for example, the definition of the measurement location on the piece, or the number and arrangement of single measurements as a basis for a good average value.
- Details of the measuring value collection method for selective inspection and of the further steps of measured value processing, or guidelines for the application of analyzing software (for example, the selection of a compensating method).
- Legal warranty of adequate qualification of the personnel who conduct the measurement.

The result of inspection planning is the inspection plan.

8.5 Further Reading

- T. M. Bosch, M. Lescure: *Laser Distance Measurements* (Atlantic Books, London 1995)
- H. Czichos, T. Saito, L. Smith (Eds.): *Springer Handbook of Materials Measurement Methods* (Springer, Berlin, Heidelberg 2006)
- E. Dietrich, A. Schulze: *Statistical Procedures for Machine and Process Qualification* (ASQ Quality Press, Milwaukee 1999)
- P. F. Dunn: *Measurement and Data Analysis for Engineering and Science* (McGraw-Hill, Columbus 2004)
- H. Pham (Ed.): *Springer Handbook of Engineering Statistics* (Springer, Berlin, Heidelberg 2006)
- H. J. Hocken, R. J. Hocken: *Coordinate Measuring Machines and Systems*, 2nd ed. (CRC, Boca Raton 2009)
- S. Vardeman, J. M. Jobe: *Statistical Quality Assurance Methods for Engineers* (Wiley, New York 1999)
- G. T. Smith: *Industrial Metrology: Surfaces and Roundness* (Springer, Berlin, Heidelberg 2002)
- W. N. Sharpe, Jr. (Ed.): *Springer Handbook of Experimental Solid Mechanics* (Springer, Berlin, Heidelberg 2008)
- L. Wisweh, M. Sandau. R. Ichimiya, S. Sakamoto: Determination of measuring uncertainty and its use for quality assessment and quality control, Research Report Faculty of Engineering Nr. 47, Niigata University, Japan (1998)
- N. Zenine, S. Wengler, L. Wisweh: Polygon connections - manufacture and measurement, Proc. VIth International Scientific Conference Coordinate Measuring Technique, Scientific Bulletin of University of Bielsko-Biala. No. 10 Bielsko-Biala (2004)

9. Engineering Design

Alois Breiing, Frank Engelmann, Timothy Gutowski

The development and design of engineering systems following a methodical approach based on information from the literature [9.1–6] is a useful procedure. The guidelines for design methodology have also been applied to interdisciplinary development projects of this type, using aids such as requirements lists, the functional structure, and morphological boxes, to name just a few. During the design phase of the product development process it is important to comply with the basic design rules: *simple*, *clear*, and *safe* [9.3]. Several examples that clearly show the realization of these three criteria are included in this chapter.

9.1 Design Theory

9.1.1 Product Planning Phase

It is possible to structure technical products in individual life stages. These are often the basis for work done by the product manufacturer, but also by the product user. Examples include schedules for the development of a product or maintenance plans.

Figure 9.1 shows essential product life stages of a product in the sequence of production and the application. For examining the structures further, it is possible to subdivide the individual product life phases into steps. In practice, this provides the engineer with a tool, which allows him to categorize his activities accurately.

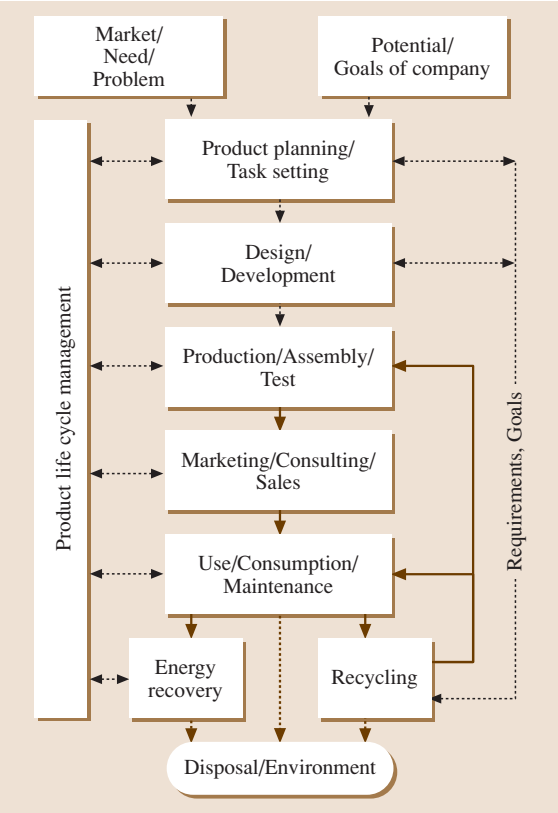


Fig. 9.1 Life cycle of a product (after [9.3])

Product Life Stages and Technical Life Cycle

The life cycle of technical products is closely linked with the general *material cycle* (Fig. 9.1). The cycle starts with an idea for a product, arising from a market or customer need. This is concretised the first phase of product life: product planning. The result is the setting of a task which provides the basis for the second phase of product life: Development and construction. At this stage, the implementation of the idea and/or task-setting into a viable product takes place in individual steps. The life cycle then continues onwards with the manufacturing process with the manufacturing of the parts, assembly and quality testing. The process ends with the product manufacturer when the product is passed on to the distribution department.

This product life phase is the interface to the application of the product, something which can be described as the product's usage or consumption. Intermediate maintenance steps can serve to extend the useful life. Product recycling follows the primary use of the product, something which can lead to a further use for the same invention or changed product functions (reuse/further use) or to secondary usage with the same or changed characteristics to the secondary materials (recycling or further utilization). Non-recyclable components will then end up in a landfill site or burnt to produce thermal energy.

Except for recycling or landfill, the life cycle applies both to the physical products of machinery, equipment and devices as well as to software products. It is usual for companies to use such structuring techniques for product tracking.

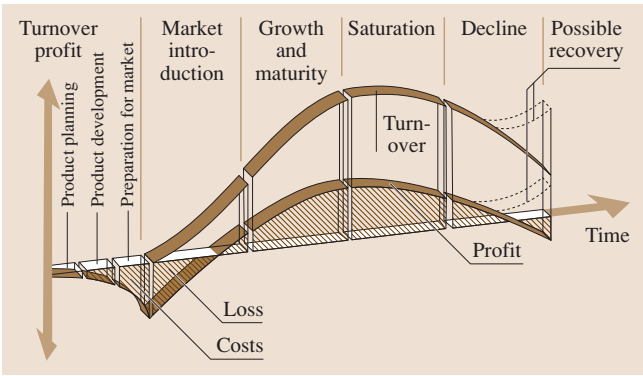


Fig. 9.2 Life cycle of a product (after [9.3])

Economic Life Cycle

The life cycle of a product can not only be seen in terms of the succession of product life stages on top of each other and/or concrete steps of manufacture and application, but also in terms of economic data, related to the respective stage of the product life. Figure 9.2 shows the relation of the product phases to turnover, profit and loss. It can be seen that before turnover starts the respective company has to make investments as implementation costs. The amount that this investment costs is heavily dependent on the product. They must first recover their investment until they break even. Only then can they gradually realize their ultimate goal to make profits.

This profit zone is characterized by a phase of growth and a phase of saturation in the market before decline sets in through a reduction in turnover and profit. A revival in sales and profits, for example, through special sales and promotional activities, is usually only for a short period, so it is more promising to achieve rising life curves of new products in a timely way through the development of new products than to offset declining life curves of old products.

Product Planning

Importance. The first two phases of product life, product planning and product development, are among the most important tasks in industry. The continuous generation of marketable products is the foundation for the economic success of the company. Because of the inevitable downturn phases for existing products or product groups (Sect. 9.1.1), the systematic planning of new products must take place, something which can also be seen as an innovative product policy [9.4]. Strategies for product planning should not be a barrier for creative companies and their engineers. Rather, these should have a supporting effect as methodological aids.

Fundamentals. The bases for the planning of products are the relationships in the market, relationships within the environment of the company and within the company itself. These can be defined as external and internal influences on a company, particularly towards its product planning.

External influences come:

- From the world economy (e.g. exchange rates)
- From the domestic economy (e.g. inflation rate, labour market situation)
- From legislative and administrative acts (e.g. environmental protection)
- From the buying market (e.g. suppliers' market and commodity market)
- From research (e.g. government-funded research priorities)
- From technology (e.g. developments in microelectronics or laser technology) as well as
- From the market

As such, the conditions of the market are crucial. A distinction can be made between a buyer's market and a seller's market. In the former, the supply is the larger than the demand and in the second the demand is larger. In a seller's market, production is the bottleneck however on the other hand, in a buyer's market, products must be designed and developed, which have to be successful in competition with the products of other providers.

Further criteria for the identification of markets are:

- Economic areas: domestic market, export markets
- New factors for the company: current market, new market
- Market position: market share, strategic free reign of the company, the technical value of its products

Internal factors come:

- From the organization of the company (e.g. product oriented vertical or task oriented horizontal organization)
- From the staff (e.g. availability of qualified development and manufacturing staff)
- From financial strength (e.g. investment opportunities)
- From the size of the company (e.g. in terms of turnover which can be sustained)
- From the production fleet (e.g. with regard to certain manufacturing technologies)
- From the product programme (e.g. with regard to components which can be adopted and predevelopments)
- From expertise (e.g. development, marketing and production experience) as well as
- From the management (e.g. as project management)

The influences listed are also described as potential of the company.

Product Development

General Approach. The second phase of product life is development and construction. This is also often referred to as product development. To further structure this phase of product life, it is usual to break stages down into individual steps. This procedural approach in handling constructive tasks is based on general solution methods and/or working method approaches as well as the general relationships in building technical products. It is not a rigidly prescribed approach, but instead, it is an essential tool for the engineer in product development. The individual working steps are the basis for other activities, e.g. the preparation of schedules or the planning of product development costs. They also help the engineer in finding where he is in the development process. A possible structure can be seen in Fig. 9.3.

Despite the variety of product developments, it is possible to work out a sector-independent flowchart, the work steps of which have to be modified to the special conditions in stetting the tasks. The approach begins with clarifying and specifying the task, something which is especially important for new design tasks. The basis for this is the stetting of tasks with individual needs which are developed from product planning tasks. From the wealth of specified requirements, the designer engineer must identify the essential problems to be solved and formulate these in the language of his field of design. The result is a requirements list, which is also known as a specification sheet. It is the

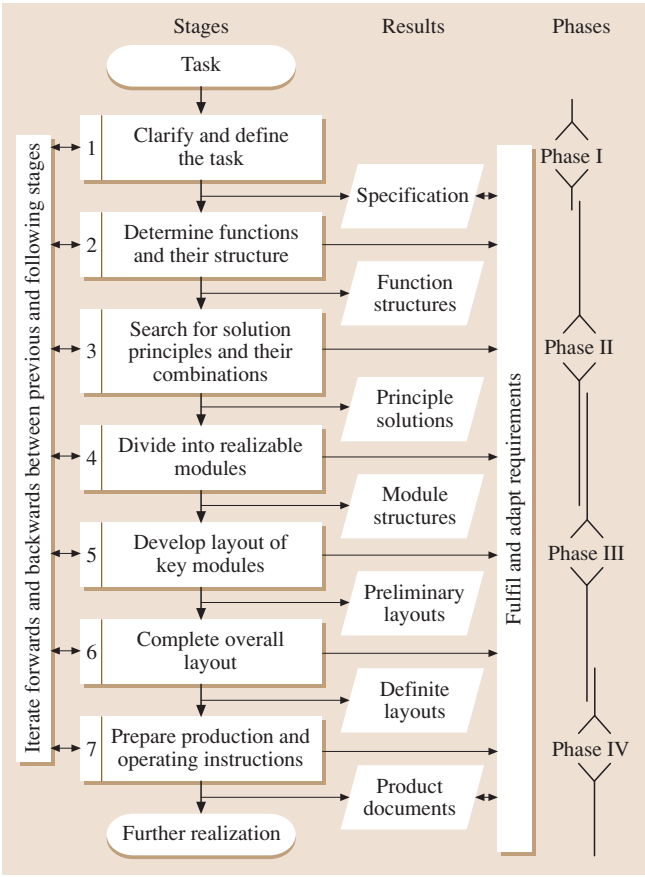


Fig. 9.3 General approach to design (after [9.3])

not only the technical but also, the legal basis for all other activities.

In the next work step, the *solution-neutral* definition of the task takes place, i. e. adopting the assumption that solutions are not prefixed. This has been proven in the form of functions, whose links lead to functional structures.

As a result of this work step, such functional structures already demonstrate an abstract form of a solution concept and as such, must be subsequently implemented gradually (Sect. 9.1.2).

The search then follows for solution principles for the key subfunctions. For mechanical products, these are based on physical effects and their fundamental realization with the help of geometric and material characteristics (Sect. 9.1.2).

The individual solution principles are represented with the help of a *morphological box*. At the same time, for each key subfunction, a maximum of three to four solution principles should be worked out. With the help of the morphological box (Fig. 9.18), the linking of individual solution principles leads to an effective structure. Even here, it is not usual to work out more than three effective structures. The effective structures which are generated in this way (or principle solutions) are compared with each other through appropriate evaluation mechanisms are (Sect. 9.1.3). Thus, it is possible, as a result of this work, to give a green light to a principle solution (also known as a concept) for further treatment. In further treatment, the principle solution is divided into realistic modules which lead to a structure and allows functional design or design priorities to take shape before the labour-intensive concretisation stage. Furthermore, it is necessary to consider the possibility of realizing a structure for further work steps which is production-like, easy to assemble, easy to maintain and recycle and/or, which is *building block-like*. The result is the modular structure.

The designing of relevant structural modules takes place in the next work step, e.g. in mechanical systems, the assemblies, components and the necessary connections are specified. Essentially, this work step includes to the following activities: Procedural calculations, stress and deformation analyses, arrangement and design considerations, as well as manufacturing and assembly examinations. As a rule, these procedures do not yet serve production-related and material engineering-related detail specifications, but first of all, the specification of key characteristics of the design structure in order to be able to optimise these according to technical and economic considerations. The results are draft concepts.

The next work step includes the design of other, usually dependent functional elements, i. e. micro-designing all sub-assemblies and components, and their combination in the overall design. For this a variety of calculation and selection methods, material catalogues, machine parts, norm parts and purchased parts are used as well as costing procedures for finding out costs. The result is the overall design.

The last work step serves to prepare tasks related to execution and usage, i. e. the development of full diagrammatic documents, including parts lists for manufacturing and assembly as well as the development of operating manuals and maintenance specifications. The result is the complete product documentation.

In practice, several work steps in the development and construction phases are often combined, e.g. for organizational or work related reasons. Thus, in me-

chanical engineering, the first three sections are seen as a *concept* phase, the next three sections as a *draft stage* and the final section is seen as a *preparation* phase.

Product Specification. The general approach has to be modified in task setting and/or for products, in which *several specialist units* are involved.

The goal is to carry out the corresponding technical tasks largely independently, but in a coordinated way. Such relationships apply, for example, to products in the biomedical field, where the medicine, biology, mechanical engineering, electrical engineering and computer science are all involved. Furthermore, precision engineering can be added to this list. Here, the design of the mechanical component, the development of the electrical and electronic circuits and control parts and the development of the necessary software take place largely independent of the different specialists.

Whilst the preparation of the requirement specifications and the functional structure takes place advantageously for the overall product, the other working steps are split into the parallel development paths, naturally in close coordination with the various specialist areas. For this, it is helpful after major concretisation jumps (e.g. to set the modular structure and after presenting the individual working drafts) to summarily document the results of the work (system structure, system draft) in order to detect missing coordination and to obtain an overall homogeneous product. Product documentation takes place for the overall product. While in the case of *new developments*, all work sections have to be followed through, steps 2 and 3 are often dispensed with in *further developments* (Fig. 9.3). The same applies in *adaptation constructions* (where steps 4 and 5 are dispensed with).

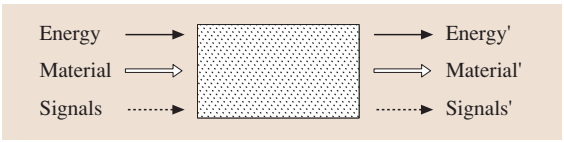


Fig. 9.5 The conversion of energy, material and signals. Solution not yet known; task or function described on the basis of inputs and outputs (after [9.3])

In many cases, it has proved advantageous, to supervise and/or reconstruct development steps again in order to make a comparison with the current state of knowledge. The development represented takes places for *products made individually*, usually only once. In the case of unsatisfactory work results, individual work sections are performed again. For *mass production* products, such as motor vehicles or household appliances, direct realization in the end product would be associated with high economic risk. According to Fig. 9.4, it is usual for such products to undergo the development and manufacturing cycle several times in order to identify vulnerable areas in the production process, initially in functional and laboratory samples and where necessary, in additional prototypes and/or pilot runs in the intermediate trial and testing phases. These are then optimized in a re-design and manufacturing process.

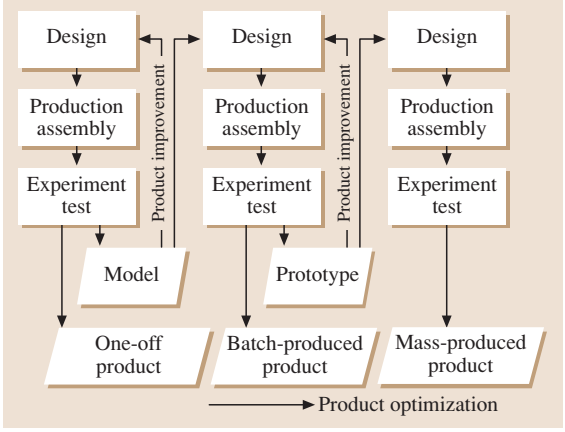


Fig. 9.4 Stepwise development of a mass-produced product (after [9.3])

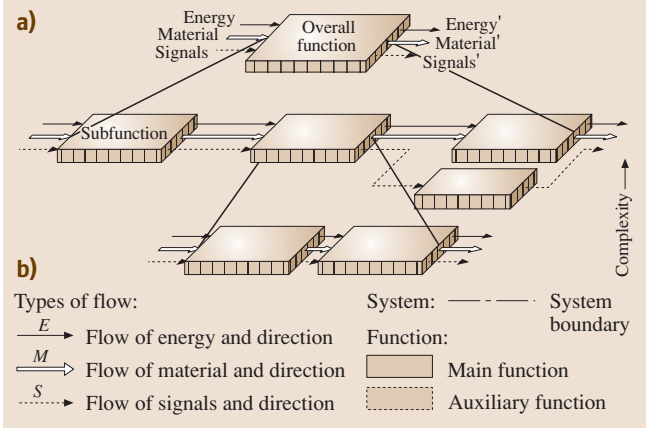


Fig. 9.6a,b Establishing a function structure by breaking down an overall function **(a)** into subfunctions **(b)** (after [9.3])

9.1.2 The Development of Technical Products

The construction of technical products is characterised by several general relationships, which also determine the different levels of specific product development.

Function Relationships

The term function, describes the general relationship between the input and output of a system with the goal of fulfilling a task (Fig. 9.7).

In the case of technical products or systems, the input and output variables are *energy variables*, *material variables* and/or *signal variables*. As a signal is the physical realization information transmission, *information* is often also selected an as input and output variable for the signal.

The target function, or target functions are an abstract, solution neutral and clear form of task setting. They result in the development of new products from requirement schedules.

Corresponding to Fig. 9.6 a difference is made between:

- The overall function for describing an overall task of a product and/or system which is to be solved.
- Subfunctions which come about by subdividing an overall function with the goal of making tasks easier to solve. At the same time, the most practical degree of subdivision is dependent on the degree to which the task setting is new, the complexity of the product to be developed as well as knowledge about solutions for fulfilling the functions.

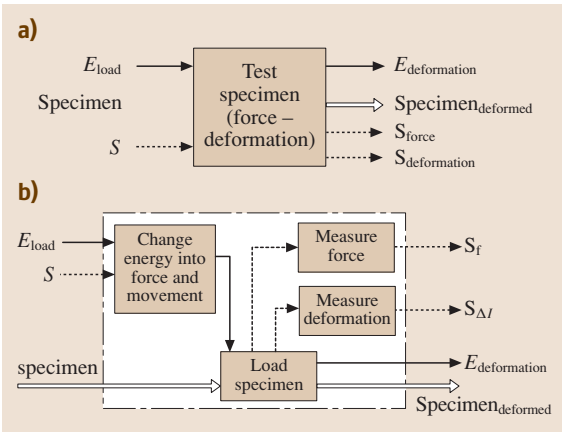


Fig. 9.7a,b Overall function (a) and subfunctions (main functions) (b) of a tensile testing machine (after [9.3])

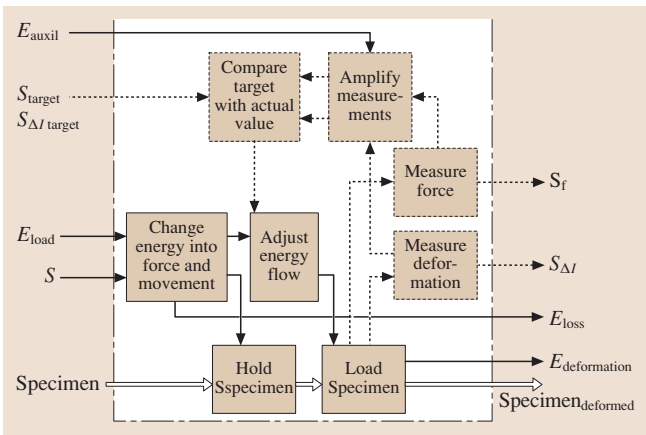


Fig. 9.8 Completed function structure for the overall function set out in Fig. 9.7 (after [9.3])

Subfunctions are linked to a *function structure*, wherein the linkages are determined by logical and/or physical compatibilities.

As an example, Figs. 9.7 and 9.8 show the function structure of a testing machine [9.3]. The overall function to be realized (overall task) is to define the stress applied to a test specimen and to measure deformity. The transfer from a function's/function structure's input to output and/or the processing of energy, material and signal variables is portrayed as energy volume, material volume, signal flow or volume. The various flows or volumes typically occur simultaneously, wherein one or several flows or volumes dominate, i. e. can determine the product. The latter are termed *main flows* or *main volumes*. They immediately serve to fulfil the function of a product. As such, we can imagine a conveyor system through which material volume as seen as the product-determining main volume, while energy volume is the driving function and the signal volume realises the control functions.

Such volumes and/or flows accompanying the main volume serve are supportive and are only indirectly related to the functional performance of the product, because they are not directly derived from the nominal functions (main functions) in the task setting. The bases for the accompanying flows are always the generated solutions for the main functions. Accordingly, they are also referred to as *tributaries* and the participating subfunctions as a *side functions*. In the function structure illustrated in Fig. 9.7, all of the listed subfunctions which result from the overall function (overall task) are main features. The subfunctions which result from the

measurement principles as necessary (*amplify measurements* and *compare target with actual value*) are, on the other hand, side functions, Fig. 9.8.

In summary there is no material or signal flow without an accompanying flow of energy, even if the energy required is very small and can easily be provided. Signal volume without accompanying material volume, e.g. in measuring devices, is however possible. Even energy volume e.g. for the production of electrical energy, is connected to material volume, wherein for controlling, the accompanying signal flow is a major tributary.

Specific Functions.

Logical Function. Bivalent or *binary* role variables often play a role when designing and describing technical systems: Conditions (fulfilled – not fulfilled), statements (true – false) and switch positions (on – off). The design of systems to realize the required dependencies between binary variables is known as logical design. It uses mathematical statement logic in the form of boolean algebra with the fundamental linkages AND and OR, and negation [9.3].

Using Boolean combination elements, complex circuits can be build which for example, increase the safety of control and reporting systems.

As an example, Fig. 9.9 shows the monitoring of a oil supply system for bearings in which respectively the nominal and actual values, the pressure monitor and the flow monitor are linked by an AND function, while the output signals of the pressure and flow monitors are connected to each other by an OR function. All bearings are linked in turn by AND links, i. e., all bearings must at least have effective oil monitoring for the machine to be ready.

Generally Valid Functions. These are increasingly recurrent functions in technical products, which can serve as sort keys for solution catalogues, as a basis for functional structural variations and as an abstraction aid in the analysis of existing products according to their functional relationships.

In Fig. 9.10, five such functions are brought together, which with the help of a cluster variation, have been derived from the input and output of a function in terms of type, variable, number, place and time. For other suggestions concerning general functions [9.3]. In this context it must be noted that such an approach and task structure is very abstract. For this reason, they are usually only used for new construction designs.

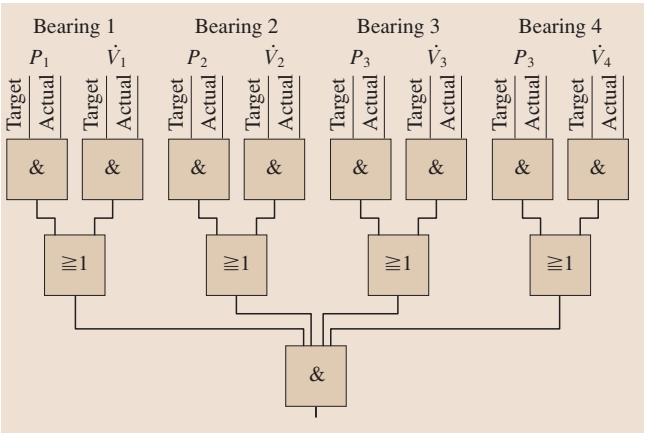


Fig. 9.9 Logical functions for monitoring a bearing lubrication system. A positive signal for every bearing (oil present) permits operation. Monitor pressure p; monitor oil flow (after [9.3])

Cause-Effect Relationships

The subfunctions and the functional structure of a technical functional relationship must be fulfilled by a cause-effect relationship. Accordingly, this arises from *active principles* for fulfilling the subfunctions and from an *active structure* for fulfilling the function structure. The *active structure* then, results from a linkage of several *active principles*. An *active principle* is determined by a physical, chemical or biological effect, by a combination of several effects as well as through their principal realisation using geometric and mater-

Characteristic Input (I) / Output (O)	Generally valid functions	Symbols	Explanations
Type	Change		Type and outward form of I and O differ
Magnitude	Vary		I < O I > O
Number	Connect		Number of I < O Number of I > O
Place	Channel		Place of I ≠ O Place of I = O
Time	Store		Time of I ≠ O

Fig. 9.10 Generally valid functions derived from the characteristics type, magnitude, number, place and time for the conversion of energy, materials and signals (after [9.3])

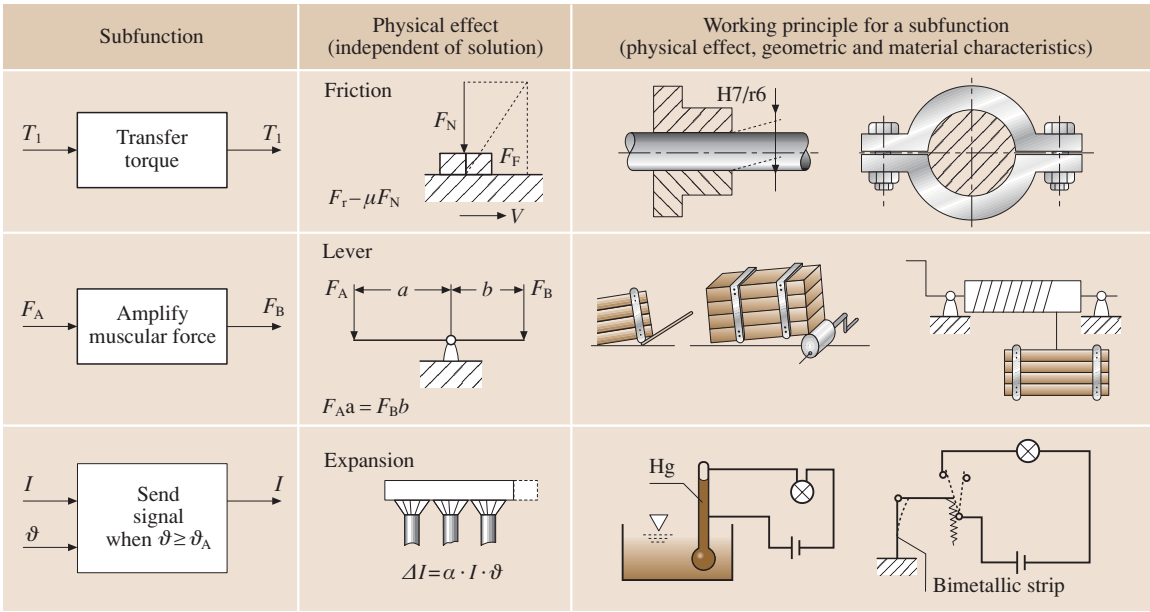


Fig. 9.11 Fulfilling subfunctions by working principles built up from physical effects and geometric and material characteristics (after [9.3])

ial characteristics (active structural characteristics). In engineering, it is usually physical effects which predominate. Examples can be seen in Fig. 9.11.

Physical, Chemical and Biological Effects. In the case of material products produced by system, machinery, equipment and manufacturing, the foundation for the solution is formed from effects, especially from physics but also from chemistry and/or biology. Effects are described by laws, which assign each of the variables involved, even quantitatively. For example, in the clutch in Fig. 9.12 the subfunction *change switching force* F_S *into normal force* F_N is realised by the physical leverage effect and the *generate tangential force* F_U is realised by the friction effect. *Rodenacker* [9.5], *Koller* [9.6] and *Roth* [9.2] in particular have all described physical effects on structures. The fulfilment of a subfunction can often only be generated by linking multiple effects which are established e.g. in the impact of a bimetal, from the effect of thermal expansion and that of Hooke's law (stress–strain relationship). The screw connection can be cited as another example. Apart from Hooke's law, here the linkage takes place between the friction effect, the wedge effect and the leverage effect.

In general, a subfunction can be fulfilled by various effects, such as the hydraulic/pneumatic effect listed in Fig. 9.12 or *force change function* which arises through the leverage effect, the wedge effect and the electromagnetic effect. From this, various solutions arise in setting tasks and with it different products with different characteristics.

Geometric and Material Characteristics. The place at which an effect or a combination of effects become(s) effective (active) is known as the *effective location*. Here, the fulfilment of the function is forced by applying the respective effect through *effective geometry*, i. e. the arrangement of effective surfaces or active areas and by selecting working motions (in the case of moving systems). With the *effective geometry*, the effective material properties already need to have been established in order for the effective relationship to be identified. Only the combination of effects and geometric/material characteristics (effective geometry, working motion and material) form the solution principle. This relationship is referred to as the active principle. The combination of several active principles leads to an (effective structure) solution (also known as the solution principle). In Fig. 9.12, the participating effective surfaces are in the form of clutch discs (friction discs) for example, and the rotational working motion of the lever to produce contact pressure can also be seen. Here, examples can be seen for the various effective surfaces, e.g. a friction clutch [9.7].

Structural Relationships

The design-related concretisation of the effective relationship leads to the *structure*. This materialises an effective structure through individual components, units and connections (Fig. 9.12) which are especially defined according to the needs of design, manufacturing, assembly and transportation with the help of scientific principles from the laws of material strengths, materials engineering, thermodynamics, fluid mechanics and manufacturing among other things. Good machine elements are also important fundamentals [9.8].

System Relationships

Technical products are components of superior systems which can be formed by persons, other technical systems and the surroundings (Fig. 9.12). At the same time, a *system* is determined by system elements and subsystems which are bordered by a system boundary. These are linked to each other and to the surroundings by energy, material and/or signal variables. A system and/or product is initially characterized by its own *system structure*. Figure 9.13 shows such a system structure for the gear out from Fig. 9.12 in combination with an elastic coupling (rotationally elastic). In a higher order system, this forms the purposeful effect (target function). Added to this are disturbing (effects) from the surroundings, side effects on the outside world and within the system as well as the effects of man and retroactive effects on human beings (Fig. 9.14). All effects must be viewed in relation to one another (system context, Fig. 9.12).

General Objectives of Technical Products

Objectives and restrictions for technical products are first received as requirements, requests and conditions in the requirement list (tasks setting) as a basis part of specific product development. However, beyond this, general objectives can be mentioned which although having a different weighting in a particular instance, have general validity. Such goals serve as a guideline for setting up requirement lists, and for the choice of solutions in the various concretisation stages of the construction process.

Table 9.1 contains a list of such general objectives for tangible products, oriented on the life stages of a product (Fig. 9.1).

Applications

The general relationships which determine the construction of technical products are important bases for several applications.

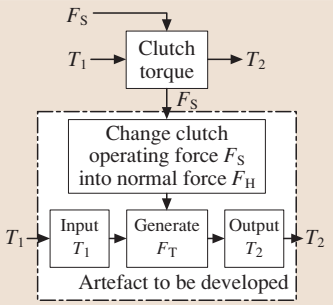
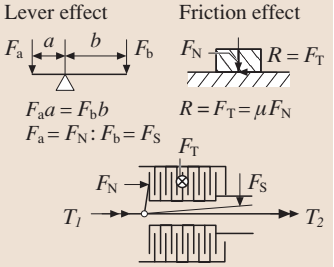
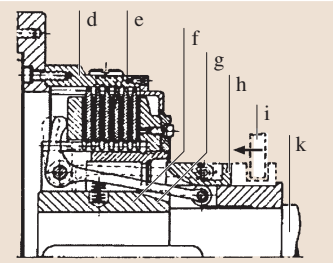
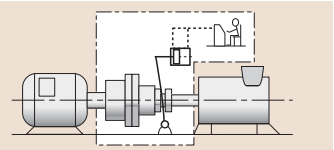
Inter-relation-ships	Elements	Structures	Examples
Functional inter-relation-ship	Functions	Functions structure	
Working inter-relation-ship	Physical effects and geometric and material characteristics → Working principles	Working structure	
Constructional inter-relation-ship	Components Joints Assemblies	Construction structure	
System inter-relation-ship	Artefacts Human beings Environment	System structure	

Fig. 9.12 Interrelationships in technical systems (after [9.3])

In product development, they enable a gradual approach, in which at first, principle solutions are sought for the required set of functions. These solutions are then concretised in design and material specifications. At each practical level, a variety of solutions can be set up as the basis for solution optimization through the different characteristics/solution characteristics of the respective relationship.

Another important area of application is the analysis of existing technical products with the aim of making an improvement, further development or adaptation to

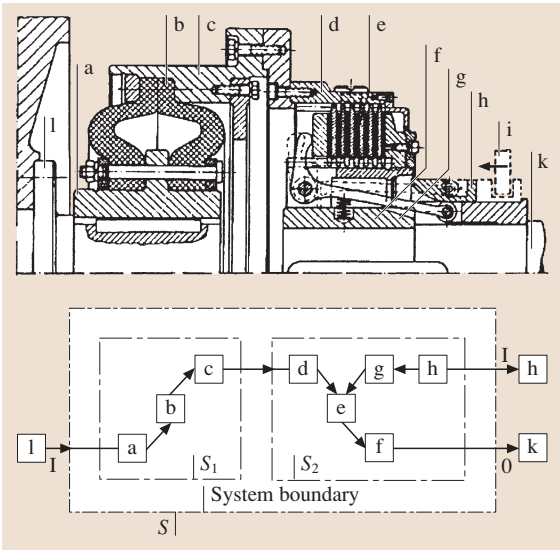


Fig. 9.13 System *coupling*: a–h system elements; i–l connecting elements, S overall system; S_1 subsystem *flexible coupling*; S_2 subsystem *clutch*; I inputs; O outputs (after [9.3])

specific conditions [9.9]. For such system analyses, procedural steps and characteristics are necessary which can be derived the general relationships. An important example which can be cited is value analysis which attempts to minimize to the functional cost of technical products [9.10]. Distinctive product features, also called *component parameters* [9.11, 12], are responsible for regulating construction catalogues and databases as well as providing useful assistance when looking

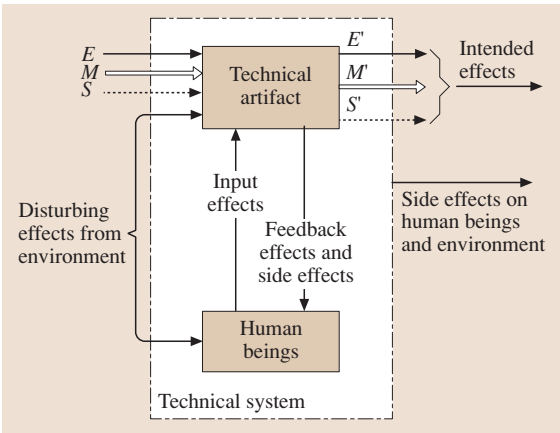


Fig. 9.14 Interrelationships in technical systems including human beings (after [9.3])

Table 9.1 General objectives for technical products (after [9.3])

Fulfil *function*
Guarantee *safety*
Take into account *ergonomics*
Simplify *manufacturing*
Make *assembly* easier, ensure *quality*
Enable *transportation*
Improve *use*
Support *maintenance*
Aim to *recycle*, minimize *costs*

for stored solutions and data from similar information storage media [9.13]. General relationships and general objectives have also been relied upon in the derivation of *component parameters* [9.14].

9.1.3 Construction Methods

General Solution Methods

Regardless of the specific degree of concretisation in the course of searching for a solution, several general methods are introduced which can also be seen as a working methodology [9.15–17]. The requirements for a methodical approach are:

- Definition of goals
- Identification of conditions
- Resolution of prejudices
- Search for variations
- Passing of judgements and
- Decisions.

General Solution Process. The solving of tasks consists of an analysis and a subsequent synthesis. This takes place in alternating work and decision steps. At the same time, it is usual that an advance is made from quantitative to qualitative steps or from abstract to concrete steps. The breakdown into work and decision steps ensures that the necessary uniformity of objectives, planning, implementation and control is maintained.

With reference to [9.18, 19], Fig. 9.15 shows a basic pattern of a general solution process. Each task setting initially effects a confrontation with an *unknown* which by gaining additional information, can be more or less resolved. A subsequent definition of the key problems to be solved specifies the task setting without predetermining solutions, thus leaving possible solu-

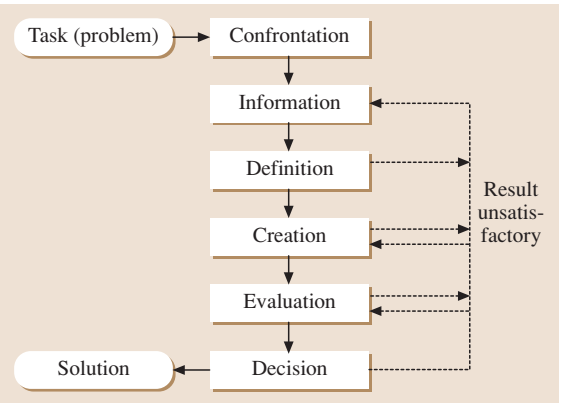


Fig. 9.15 General problem solving process (after [9.3])

tions open. The subsequent creative phase involves the actual finding of solutions. Where there are several appropriate ways, an evaluation of these has to be made in order to be able to make a decision concerning the best solution. In the case of a work step delivering unsatisfactory results, it or several work steps must be repeated, wherein in repeating the work process, even better results can be expected through a higher level of information. Therefore, this iterative process can be interpreted as a learning process.

Systematic Approach. As an inter-disciplinary science, system technology has developed methods for analysis, planning, selection and the optimum design of complex systems [9.20]. Based on system definition, a process model is introduced which is used for the different life stages of a system (Fig. 9.1). It can be seen in Fig. 9.16 that the work steps are practically identical with those in Fig. 9.15 and that the time continuum in a system runs from the abstract to the concrete.

Problem and System Structure. New and complex tasks are usually easier to solve if the overall problem to be solved is initially divided up into subproblems and individual problems. This way, subsolutions or individual solutions can be found [9.21].

A methodological basis for this approach is the structuring of systems into subsystems and system elements for the easier identification of relationships and effects within the system and outwards, on the surroundings. The extent to which a system is broken down is determined by utilitarian considerations and depends on the extent to which the problem is new and the knowledge of the engineer/scientist.

Such structuring also encourages the adoption of known and proven subsolutions, the working out of alternative solutions, a systematization technique for using solution catalogues and databases, the identification of integrated relationships as well as the introduction of rational working divisions.

While the overall division of problems into individual problems makes solutions easier to find, the subsequent combination process, of linking the subsolutions to the overall solution creates problems concerning the compatibility of subsolutions among themselves. Combination schemes such as that described by *Zwicky* [9.22] e.g. the morphological box, (Fig. 9.17) have proved an important tool which in a two-dimensional order scheme, assigns subsolutions to the subfunctions to be fulfilled.

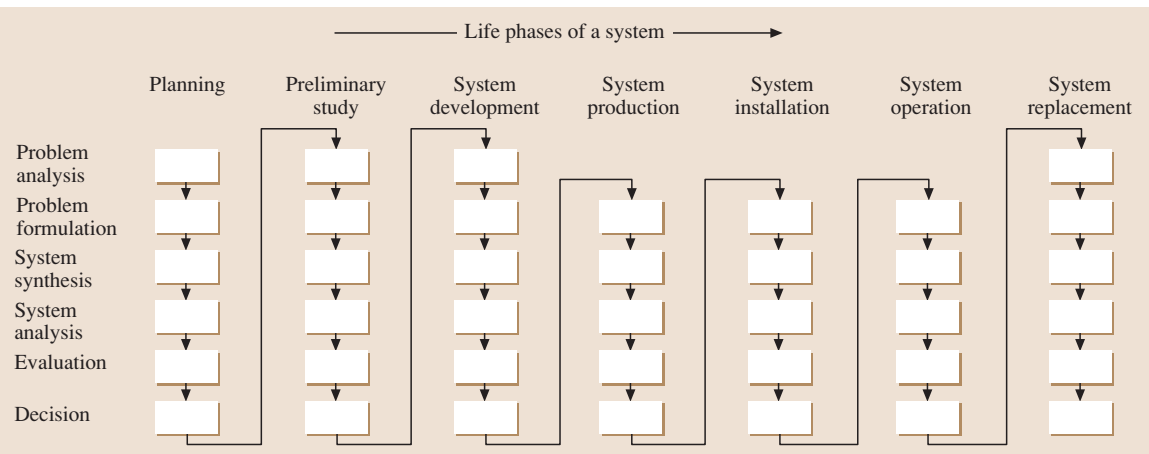


Fig. 9.16 Model of the systems approach (after [9.3])

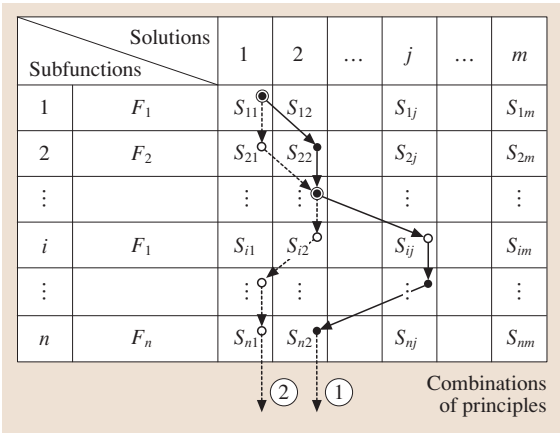


Fig. 9.17 Combining solution principles into combinations of principles: Combination 1: $S_{11} + S_{22} + \ldots S_{n2}$; Combination 2: $S_{11} + S_{21} \ldots S_{n1}$

There are also task settings for which a problem breakdown at the beginning of the solution process would not be helpful, but instead, tasks which require the initial elaboration of an integrated approach. Products for which the *industrial design* has a special significance are typical examples of this, e.g. cars or household appliances. Here, the conception of the overall appearance including ergonomic features takes a higher priority than construction details [9.23]. Industrial design and methodological problem solving are not opposites. Rather, in this case, the methodological breakdown of problems and the finding of solutions are only applied after adopting a draft for the overall appearance of the product.

General Aids. *Literature research* in books, journals, patents and company documents provide a summary of the state of the art and the competitors. They also offer the solution-searching designer new suggestions.

Through the *analysis of natural systems* (bionics, biomechanics) it is possible to identify natural forms, structures, organisms and processes and use these principles for technical solutions. For the designer, nature can provide many suggestions [9.24–27].

Through the *analysis of known technical systems*, whether those of a scientist's/designer's own company or whether those of competitors it is possible to transfer trusted solutions to new task settings and to identify fruitful further developments or alternative solutions [9.3].

Analogical considerations enable the transfer of problem to be solved or system to be identified onto an analogous system, solved problem or identified system. Ostensibly, this assists in the investigation or assessment of a system's characteristics as well as facilitates simulation or modelling [9.3].

In order to particularly identify new solution characteristics and to carry out step by step further developments, *measurements* of running systems and *model experiments* taking advantage of similar mechanics are among the main sources of information for designers. *Heuristic operations* increase creativity in the search for solutions, primarily in conventional approaches carried out by people. These operations are also known as creativity technology and are understood as tool for methodological solving and as an introduction for thinking and working in an orderly and effective form. They also repeatedly appear alongside special solution and approach methods [9.17].

Conception Methods

In conceiving working out a fundamental solution principle or a solution concept for a task setting (function) (Sect. 9.1.3), the following methods are suitable for locating principle solutions. They can be used of course, in each individual case, for more concrete design tasks.

Intuitive Methods. Intuitively stressed methods exploit group dynamic effects with which the intuition of people be stimulated through mutual association between the partners. At the same time, intuition is understood as being an imagination-stressed, barely influenceable or comprehensible action which produces ideas for solutions from the subconscious or conscious: something which can be termed *primary creativity* [9.28]. In [9.3], the following methods shall be described in detail: During the dialogue method discuss two equal partners discuss the solution to a problem, wherein as a rule, a first solution approach is assumed.

With brainstorming, a group meeting with an interdisciplinary composition takes place without aids. Ideas should be expressed without criticism and evaluation *quantity goes before quality*.

With synetics, during the group meeting, additional analogies from non-technical areas or semi-technical areas are used to generate ideas.

Method 635 is a brainwriting method in which 6 participants each express 3 problem solving ideas, in written form in 5 rounds. The proposals of the previous round are known to the participants and the level of information is constantly increased. The *gallery method* combines individual work with group work in such a way that each individually worked out proposals in the

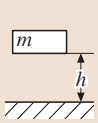
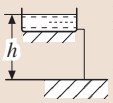
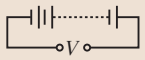
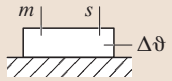
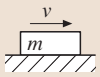
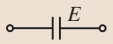
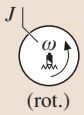
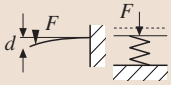
Type of energy Working principle	Mechanical	Hydraulic	Electrical	Thermal
1	 Potential energy	 Liquid reservoir (potential energy)	Battery 	Mass 
2	 Moving mass (transl.)	Flowing liquid	Capacitor (electr. field) 	Heated liquid
3	 Flywheel (rot.)			Superheated steam
4	 Wheel on inclined plane (rot. + transl. + pot.)			
5	Metal spring 	Other springs (compression of fluid + gas) 		
6		 Hydraulic reservoir a. Blader b. Piston c. Membrane (pressure energy)		

Fig. 9.18 Different working principles that satisfy the *store energy* function obtained by varying the type of energy (after [9.3])

form of group sketches are hung up in a kind of gallery. Through the discussion of these suggestions with appropriate connotations, other solutions or improvements are hit upon which in turn are worked out by the group members individually. Evaluation and selection again takes place in a group meeting.

Discursive Methods. Discursive stressed methods consciously seek solutions through gradual, influenceable and documentable procedures (*secondary creativity* [9.28]).

In the *systematic study of the physical events* are derived from a known physical relationship (a physical effect), with several physical variables are derived by different solutions, that analyse the relationship between a dependent and an independent variable with all other variables held constant. Another possibility lies in dissecting known physical effects into individual effects and to look for the implementation for these effects [9.5]. A *systematic search with the help of order schemes* assumes that an order scheme (for example, a two-dimensional table) stimulates the search for other solutions in certain directions outside of the identification of key solution characteristics and corresponding linkage options. Possible starting points are one or more known solutions which are characterised according to

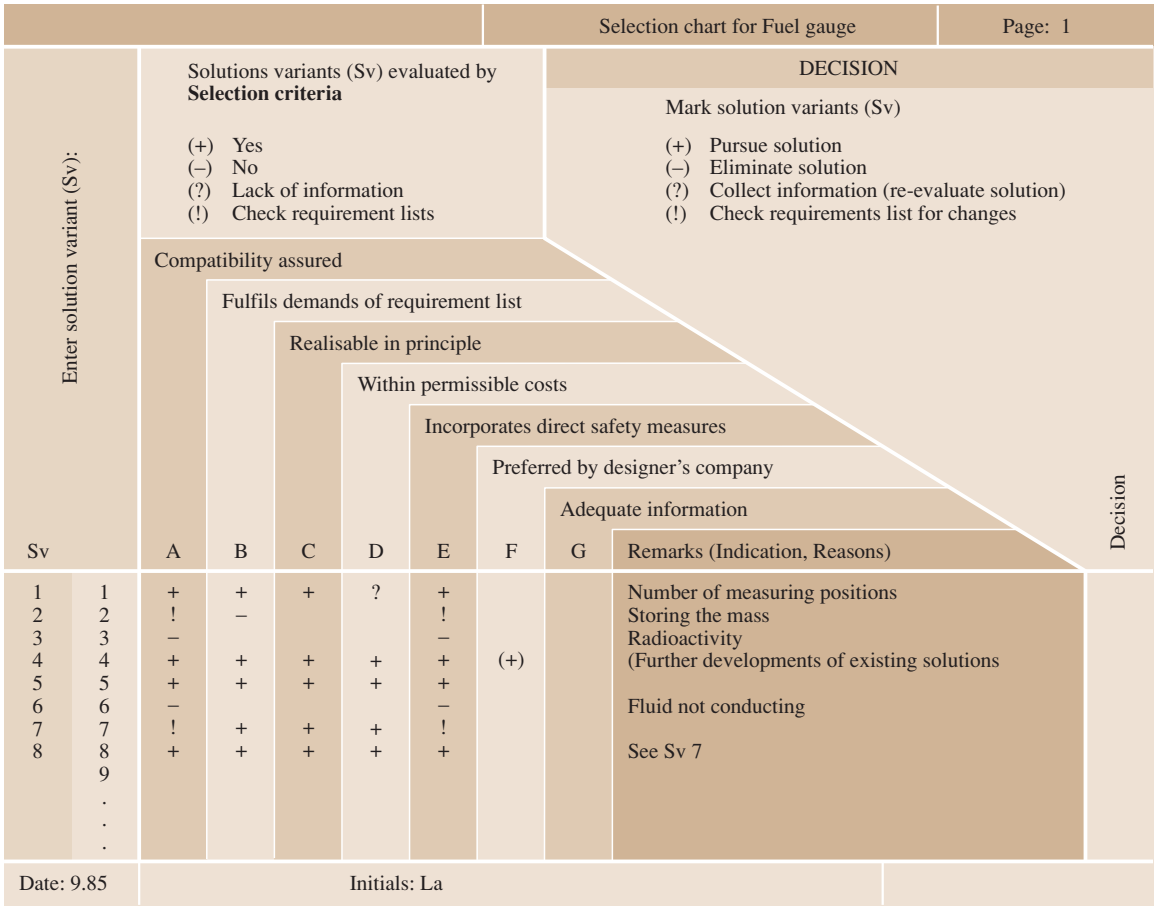


Fig. 9.19 Systematic selection chart: 1, 2, 3, *etc*. are solution variants (after [9.3])

organizational aspects or differentiating features. Such organizational aspects and/or variation features are for example, types of energy as well as the effective structural features of effective geometry, effective movement and substance. Such an order scheme is Zwicky's morphological box (Fig. 9.17).

The designer quickly arrives at solution proposals by *using construction catalogues* as collections of known and trusted solutions of various concretisation and complexity levels, but they often have to be developed or adapted [9.2]. What is important is the allocation of selection characteristics in the accessed part of a catalogue in order to recognize the suitability of a solution for the realization of a required function (task) (Fig. 9.18). Catalogues and databases are also naturally important works in looking for design options in the draft phase of product development.

Methods of Selection and Evaluation

In each design phase or concretisation stage of the development and construction process, selection methods serve to assess and select solution variations with the goal of recognizing those solution variations for which further realization is worthwhile from the quantity of solution options. Depending on the level of knowledge about the characteristics of a solution to be reviewed, procedures are employed to make a rough selection or a finer selection. A rough selection is characterized by the tasks of rejecting (–) and preferring (+). With the help of a selection list (Fig. 9.19), the initial totally unsuitable solutions can be rejected. Such rough selection processes have proved successful, especially for listing and/or designation in morphological boxes connected with the drafting of effective structures.

Table 9.2 Points awarded in the utility analysis and VDI guideline 2225 (after [9.3])

Value scale			
Use-value analysis		Guideline VDI 2225	
Pts.	Meaning	Pts.	Meaning
0	absolutely useless solution	0	unsatisfactory
1	very inadequate solution		
2	weak solution	1	just tolerable
3	tolerable solution		
4	adequate solution	2	adequate
5	satisfactory solution		
6	good solution with few drawbacks	3	good
7	good solution		
8	very good solution	4	very good (ideal)
9	solution exceeding the requirement		
10	ideal solution		

If several solutions remain, they are obviously preferred. The selection criteria are to be adjusted to the goals of product development and the company. For a more accurate selection, evaluation procedures are used, in particular, VDI guideline 2225 [9.29] and the utility analysis [9.30]. In Table 9.2, a comparison is shown between the two procedures. A detailed approach can be taken from VDI guideline 2225 [9.30].

Design Principles

After evaluating the effective structures which have been worked out and/or principle solutions, a structure/solution is usually released for drafting. The design stage in the drafting of a product requires the use of mechanics, the knowledge of strength science as well as knowledge of manufacturing technology, materials technology and other fields. The fine shape is gradually generated from the rough shape:

- Rough design: spatially and significantly correct, but without details, i. e. preliminary drafts
- Fine design: all the necessary details are conclusively defined by applying guidelines / regulations, norms, calculations and consider the impact of auxiliary functions

When generating fine shape, it is appropriate to structure the approach in individual work steps.

The starting point is the principle solution. Following clarification of the spatial conditions, the designing of the design-determining main functional elements begins and following this, the designing of the other main functional bodies. If they are sufficiently specified, the search takes place for solutions to the auxiliary functional elements [9.3]. In this step, these are often bought-in parts. The result of this working step is to define the design of the principle solution, i. e. all the characteristics of geometry, material and condition.

The following methods and rules are recommendations, strategies and hints for the designer with which he can successfully work out a structure for a product [9.3].

Basic Design Rules. Basic rules are always valid instructions, whose observance helps ensure the success of a solution and whose non-observance leads to major drawbacks. They are derived from general objectives in the construction process.

Observance of the basic rules:

- Easy
- Clear and
- Safe

leads to the clear fulfilment of the technical function, its economic realization and to safety for humans and the environment.

Observance of the basic rule *clear* helps, to reliably predict the effect and behaviour of structures. Figure 9.20 shows an example of a shaft-hub connection. This is a cross-compression connection. The additional parallel key does not make it any more secure. Sectional weakening results and there are additional notches (lo-

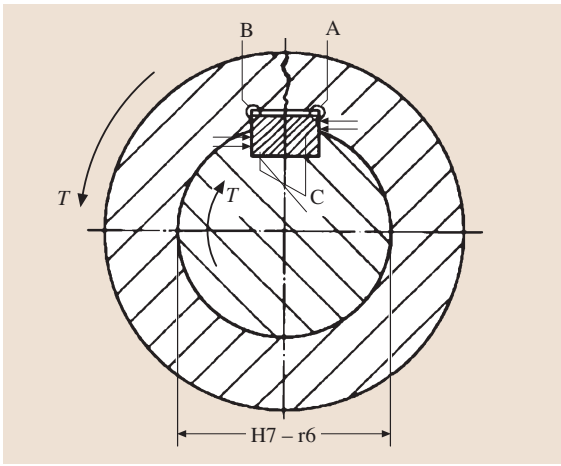


Fig. 9.20 Combined shaft–hub connection achieved by means of shrink fit and key: an example of not applying the principle of clarity

cation B) and thus, the stress condition and handling is made more complicated (location C). As a result, the whole connection is insecure!

The basic rule criterion *simple* should ensure an economic (cost effective, feasible) solution however the basis has to be formed by the functional performance. This functional fulfilment determines where the border for the *easy* criterion lies.

It is possible to apply the following basic principle: Few parts with simple design! To enforce this principle a compromise is often needs to be made between the two aspects:

- Functional fulfilment and
- Economic efficiency

The functional fulfilment conditionally requires, among other things, a minimum number of parts which must have the necessary shape to fulfil the required function.

Economic efficiency requires the necessary number of parts and the necessary shapes which emerge to be manufactured cheaply and fast (on schedule).

The demand for safety forces designers to consistently take into account durability, reliability, the extent to which the part is free from accidents as well as environmental protection. The following criteria are available to the designer at various levels [9.3]:

- Immediate safety engineering (*safe existence*, *limited failure*, *back-up arrangements*
- Indirect safety engineering (protection systems, protective equipment) and
- Indicative safety engineering (identification of the danger)

Figure 9.21 shows the key safety areas.

From the designer's point of view, the stage of immediate safety engineering is always strived for. For realizing this, three principles are available.

The principle of safe existence (safe-life behaviour) ensures that all components and their relationship in the product survive the intended stress and operational life without a failure or a fault.

The principle of limited failure (fail-safe behaviour) allows for a functional fault or damage during the operational life without at the same time causing serious consequential damage.

The *principle of back-up arrangements* enhances security through reserve elements taking over the full or partial function of the regular element in the event of failure. In the case of active back-up, both normal elements and reserve elements actively take part in fulfilling the function, however in passive back-up, the reserve element is only in reserve during normal operation. Principle redundancy exists when the normal element and reserve element depend on different effective principles. Back-up elements can be engaged in parallel, serial, quartet, cross quartet, 2 out of 3 and comparative back-up running.

If, with the three principles listed, risk can not be excluded, complementary indirect safety equipment and the indicative safety equipment is used.

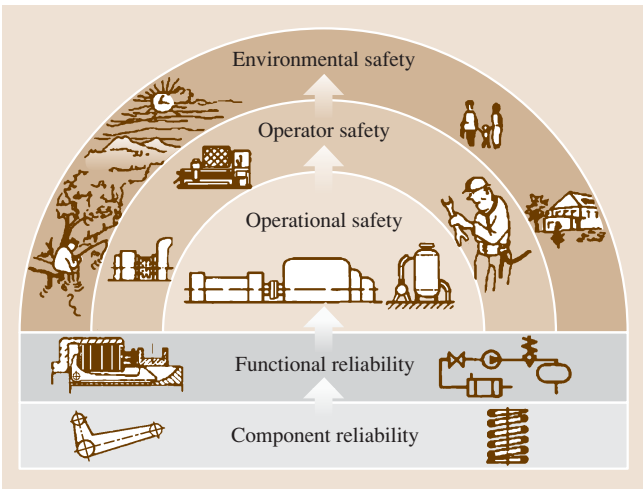


Fig. 9.21 Relationship between component and functional reliability on the one hand and operational, operator and environmental safety on the other (after [9.3])

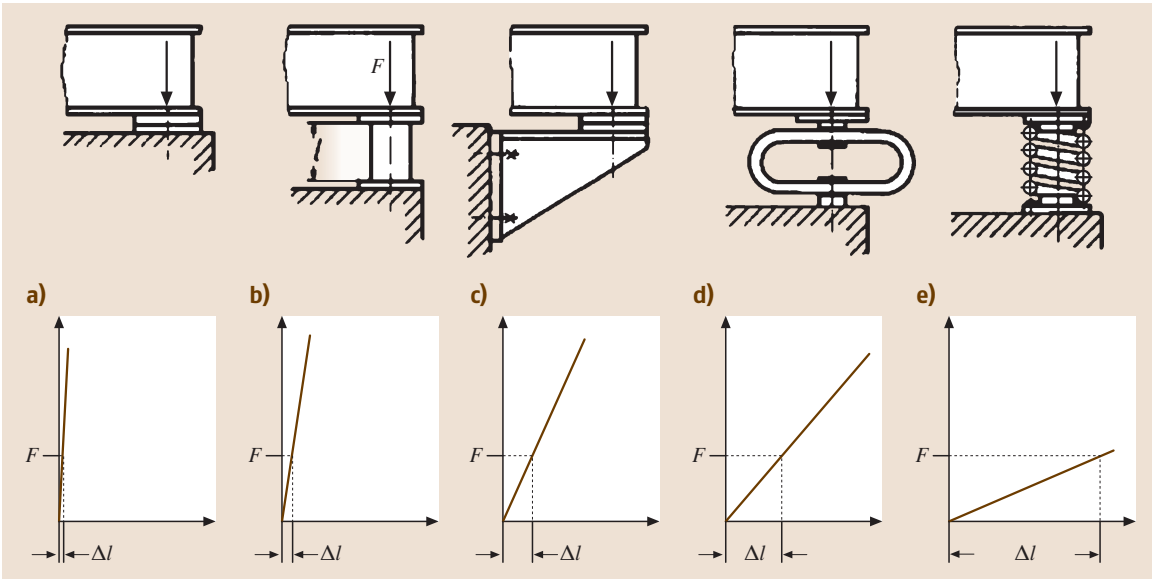


Fig. 9.22a–e Supporting a machine frame on a concrete foundation: **(a)** very rigid support due to short force transmission path and low stress on the baseplates, **(b)** longer force transmission path, but still a rigid support with tubes or box sections under compression, **(c)** less rigid support with pronounced bending deformation (a stiffer construction would involve the greater use of materials), **(d)** more flexible support under bending stresses, **(e)** very flexible support using a spring, which transmits the load in torsion. This can be used to alter the resonance characteristics (after [9.3])

Design Guidelines. In this section, the principles will be explained for developing a construction on the basis of the effective structure as well as defining the nature and (especially) the setting up of functional bodies. These principles are:

- The force transmission
- The division of tasks
- Self help
- Stability and bi-stability
- Low-fault design

Force transmission principles ensure equal shape stability, economic and load-favourable channelling of the flow of force/power flow, adjustment of component shaping as well as force equilibrium. It is worth additionally noting that the following subfunctions of force are implemented in many machine engineered products (including precision engineering):

- Pick up (induction)
- Channelling (onward channelling)
- Release (channel off)

(momentum belongs to these forces). When considering power channelling problems, the image of force flow is often very helpful. In designing the following guidelines should always be adhered to:

- The flow of force must always be closed.
- Sharp deflections in the flow of power.
- Changes in the density of power flow through strong cross-sectional changes are to be avoided.

This idea of the power flow should be supplemented through observation of the following principles:

- The *principle of equal shape stability* means that throughout the entire component, the same stable load is strived for. Economic aspects (costs) can oppose the application of this principle.
- The *principle of the direct and brief transmission of force* refers to the selection of the most direct and shortest path of channelled force (momentum) preferably with torsion/pressure stress in order to keep deformation small and material expenses low through uniform stress distribution. Figure 9.22 describes these relationships in a brace support for a machine frame on the basis of compression stress–strain characteristics of the variants. The instance when either a more rigid or elastic solution is used, depends on the design requirements.

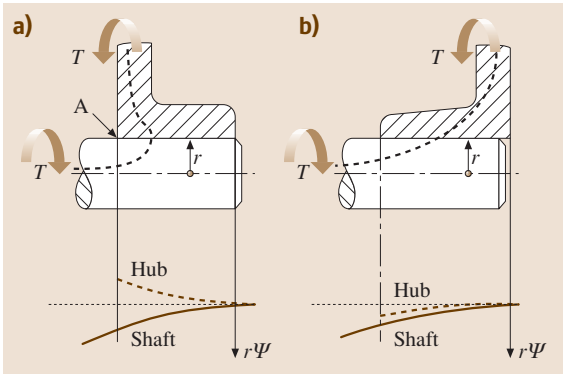


Fig. 9.23 **(a)** Shaft–hub connection with strong *force flow-line deflection*. Torsional deformations of shaft and hub in opposite direction (ψ = angle of twist). **(b)** Shaft–hub connection with gradual *force flow line deflection*. Torsional deformations of shaft and hub in the same direction (after [9.3])

- In joint connections, the *principle of adjusted deformations* designs the components concerned such that when subjected to loads, extensive adaptation of the deformation occurs, through parallel and equal deformation. Figure 9.23 shows the example of a torque-stressed shaft-hub connection (shaft to collar connection) in a favourable and less favourable design. From Fig. 9.24, it is possible to see deformation adjustment possibilities in a crane drive mechanism, without which, off-track running of one of the drives would occur. The cause for this is the different torsion stiffness of the shaft section I_1 (large) and I_2 (less than that of I_1) as shown in Fig. 9.24a. When applying torque, first the left wheel moves, the right wheel stands still and the drive remains obliquely positioned. This deficiency can be overcome through a symmetrical arrangement (Fig. 9.24b) or through the adjustment of the torsion stiffness in both shaft sections (Fig. 9.24c).
- The *principle of balancing force* aims, through compensatory elements or symmetrical arrangement, to limit the auxiliary variables accompanying the main function variables to as small an area as possible so that construction costs and energy losses remain as low as possible (Fig. 9.25).

The *principle of sharing tasks* enables the clear, safe behaviour of these functional agents, better efficiency and increased capacity by assigning components or assemblies, materials or other construction elements to individual subfunctions of a solution concept. This principle of a *differential design* stands in contrast to the usually cheaper variant, *integral design*. The usefulness of the application should be tested on an individual cases basis. By way of example, Fig. 9.25 shows a fixed bearing arrangement, in which radial forces are transferred by a roller bearing and axial forces by a deep groove ball bearing. At high loads, this arrangement is superior to the usual version with only one deep groove ball bearing, which transfers the radial and axial forces.

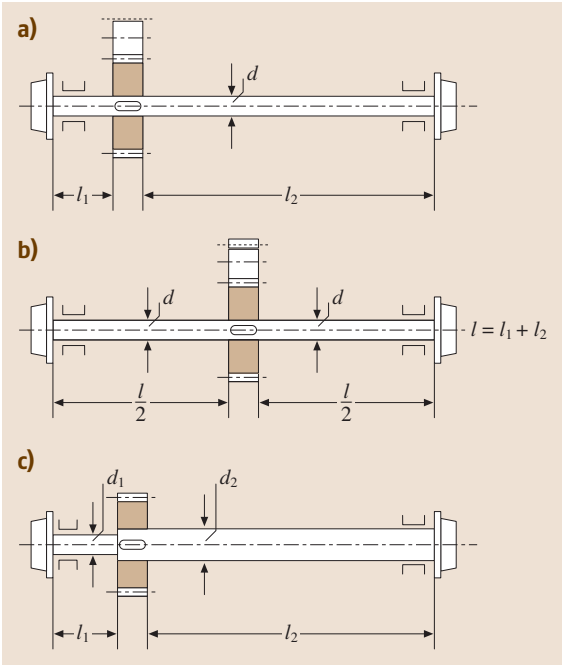


Fig. 9.24a–c Application of the principle of matched – here equal – deformations in crane drives: **(a)** unequal torsional deformation of lengths ℓ_1 and ℓ_2, **(b)** symmetrical layout ensures equal torsional deformation, **(c)** asymmetrical layout with equal torsional deformation due to adaptation of torsional stiffnesses (after [9.3])

The principle of sharing tasks is also applied for distributing load onto several identical mechanical transmission elements, if only one mechanical transmission element exceeds the load limit. Examples of this are split torque multi-channel drives and belt gear with several parallel V-belts.

Through the appropriate selection and arrangement of components to form a construction, the *principle of self-help* leads to effective mutual support, which helps a function to be fulfilled, better, safer and more economically [9.31]. At the same time, a self-strengthening and self-compensating effect can be exploited on a normal

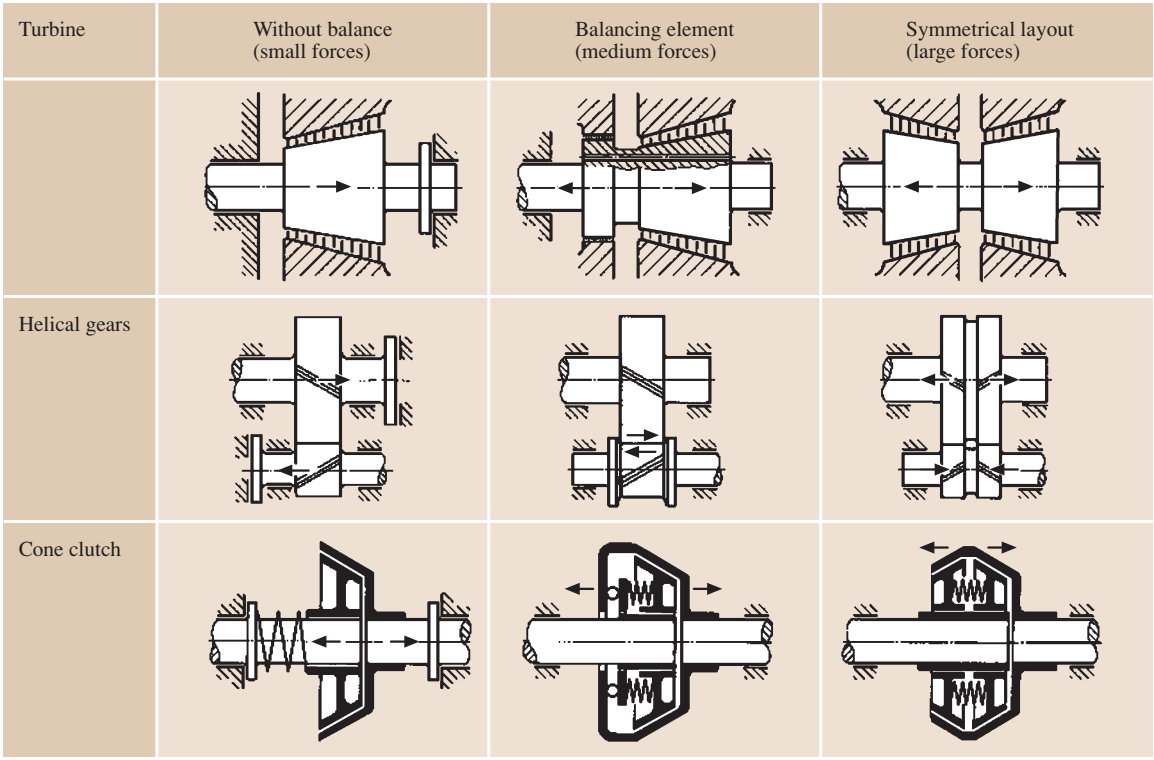


Fig. 9.25 Fundamental solutions for balancing associated forces, illustrated via a turbine, helical gears and cone clutch (after [9.3])

load as well as a self-protecting effect on an overload situation. Figure 9.27 illustrates the self-strengthening solution for a closure in pressure vessels, in which the sealing force of the lid is proportionately increased by the internal pressure of the container.

A self-compensating solution can be found in the example of an incorrectly clamped blade of a jet engine rotor. Through the slanting of the blades, additional

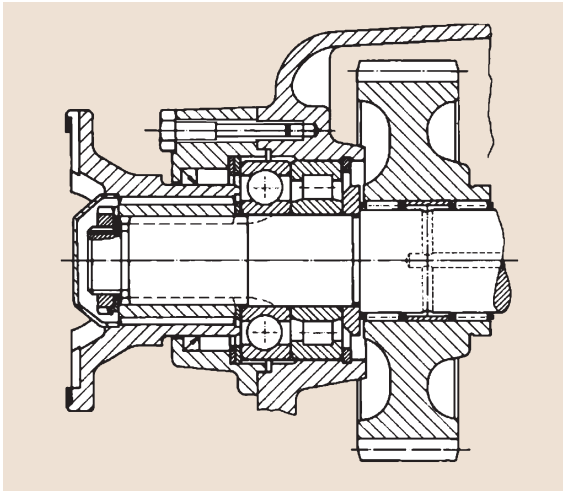


Fig. 9.26 Locating bearing with separate transmission paths for radial and axial forces (after [9.3])

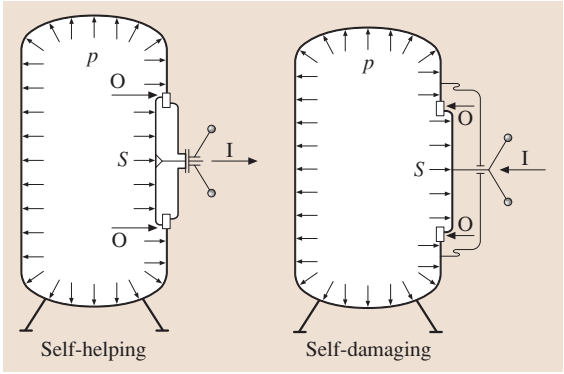


Fig. 9.27 Layout of an inspection cover. I = initial effect; O = overall effect; p = internal pressure (after [9.3])

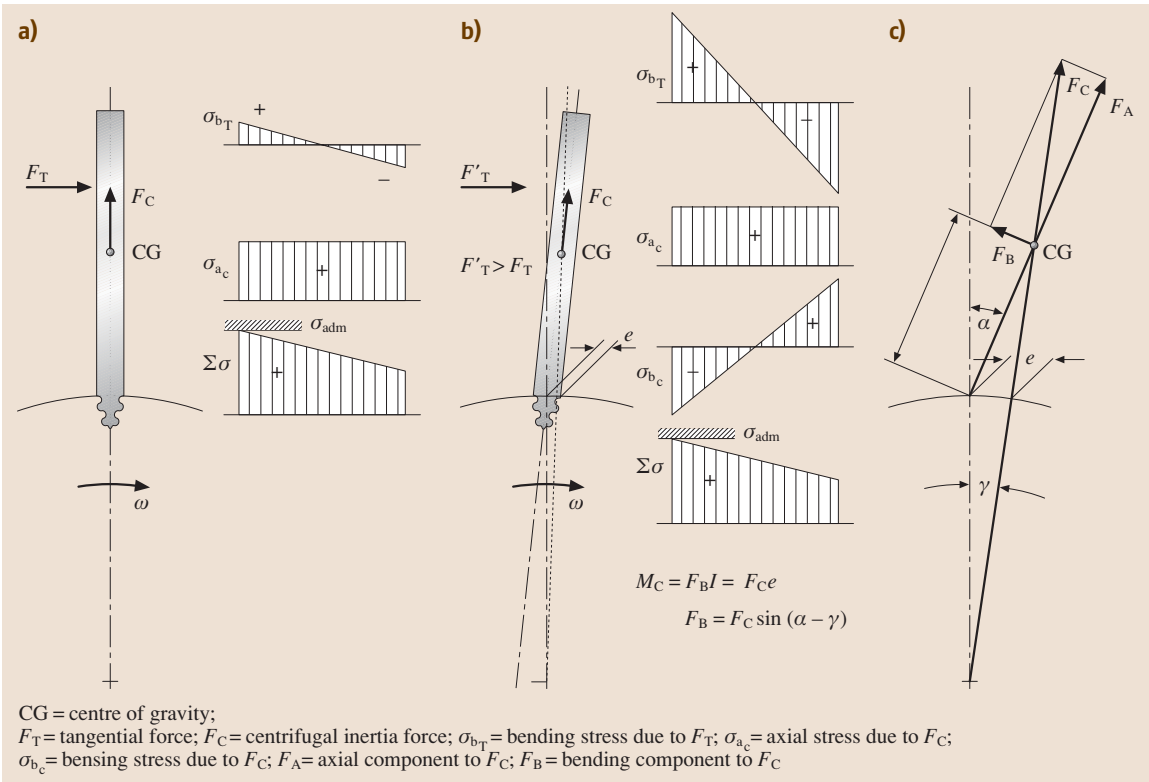


Fig. 9.28a–c Self-compensating solution for turbine blades: **(a)** conventional solution, **(b)** leaning of the blade produces a balancing supplementary effect due to the additional bending stresses produced by the centrifugal inertia force (σ_{b_C}), which oppose the bending stresses caused by the tangential force (σ_{b_T}), **(c)** diagram of forces

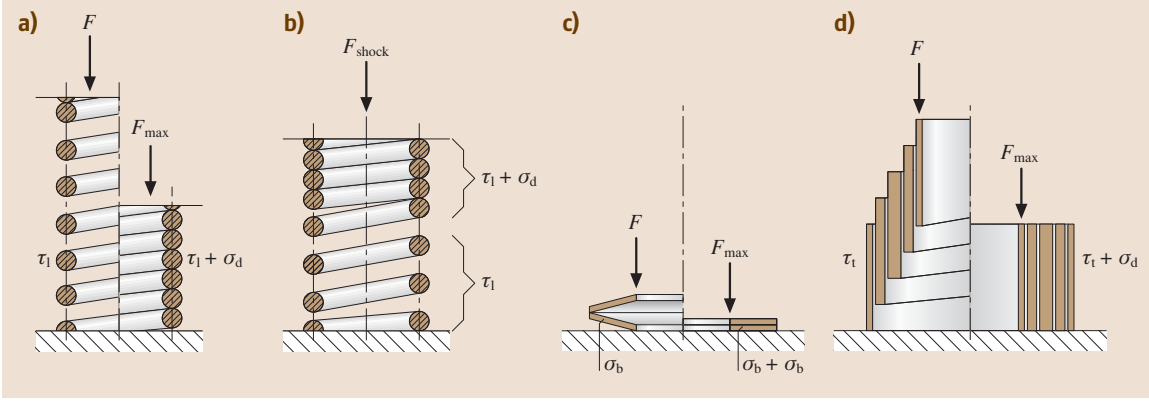


Fig. 9.29a–d Self-protecting solution in springs: **(a)**–**(d)** force transmission path changed, the normal function is suspended or limited in the case of excess loading (after [9.3])

bending stress is generated due to centrifugal force which counteracts and (partly) compensates the bending stress from the tangential force, thus enabling a greater amount of tangential force (= blade force) (Fig. 9.27).

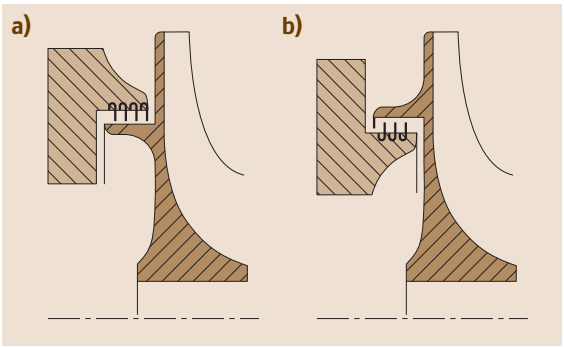


Fig. 9.30a,b Seal in turbocharger (after [9.3])

In restricting functionality, a *self-protecting* solution protects an element from overstress by making a change to the stress type (as Fig. 9.29 shows using springs as an example).

In the event of faults, the *principle of stability* has the aim of evoking a self-neutralizing compensatory effect or at least, an alleviating effect. Figure 9.30 shows this principle by means of a compensation piston seal, which when heated (fault) either starts grinding (unstable solution) or lifts itself away from the anti-effective area (stable solution). Figure 9.30a shows that the friction heat due to rubbing (fault) primarily flows into the inner part i. e. its additional warming, and the resulting expansion reinforces the fault and leads to an unstable behaviour. From Fig. 9.30b, we see that the friction heat due to rubbing (fault) primarily goes into the outer part, i. e. its additional warming and the resulting expansion acts contrary to the fault, reduces it, and leads to stable behaviour.

With the principle of bi-stability, through a deliberate fault, effects are achieved which support and thus help reinforce the fault so that when reaching a borderline state, a significantly new, clearly different state is reached without unwanted intermediate states. As such, the principle also underlies the directness of an effective structure. Figure 9.31 shows the principle of a safety valve, which should quickly move from the closed borderline state to the opened borderline state (through a sudden increase in the pressure area (A_v to A_z) after lifting the valve plate).

More Recommended Design Guidelines. The following design guidelines are recommendations for the designer, which he should note in order to satisfy the general and specific objectives of the task. A detailed description of these design guidelines is found in [9.3]. *Design in accordance with stress* means starting out

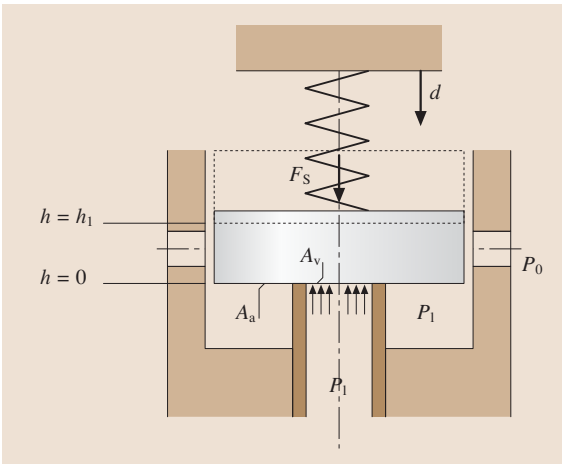


Fig. 9.31 Solution principle for a valve with an unstable opening mechanism: d = precompression of spring; s = stiffness of spring; F_S = spring force; h = lift of valve head; p = pressure on valve; p_ℓ = limiting pressure just sufficient to open the valve; p_i = intermediate pressure upon opening of valve; p' = pressure after opening of valve; p_0 = atmospheric pressure; A_v = surface area of valve opening; A_a = additional surface area.
Valve closed: $F_S = sd > pA_v$, $h = 0$;
Valve just open: $F_S = sdp_\ell A_v$, $h = 0$;
Valve opening fully: $F_S = s(d+h) < pA_v + p_i A_a$, $h = \pm h_1$;
Valve fully open: $F_S = s(d + h_1) = p'(A_v + A_a)$, $h = h_1$ (new equilibrium position) (after [9.3])

with the aim of initially ascertaining all longitudinal and transverse forces touching the component as well as bending and torque moments. This is the basis for calculating existing normal stresses, as tensile, compressed and bending stresses as well as the tangential stresses as shear stresses and torsional stresses. This stress analysis is the basis for determining the existing elastic and/or plastic deformation (strain analysis). In order to identify the level of safety against failure or to make lifetime predictions, these loads are contrasted to the applicable material threshold values for the current case load, observing notching effects as well as surface and variable influences with the help of stability hypotheses.

At the same time, the principle of *equal stability* should be strived for so that all design areas are roughly used to the same extent.

Design in accordance with expansion means thermal and tension-conditioned component expansion, especially relative expansion between the components, as compensated for by adopting channels and through the selection of materials so that no residual stresses,

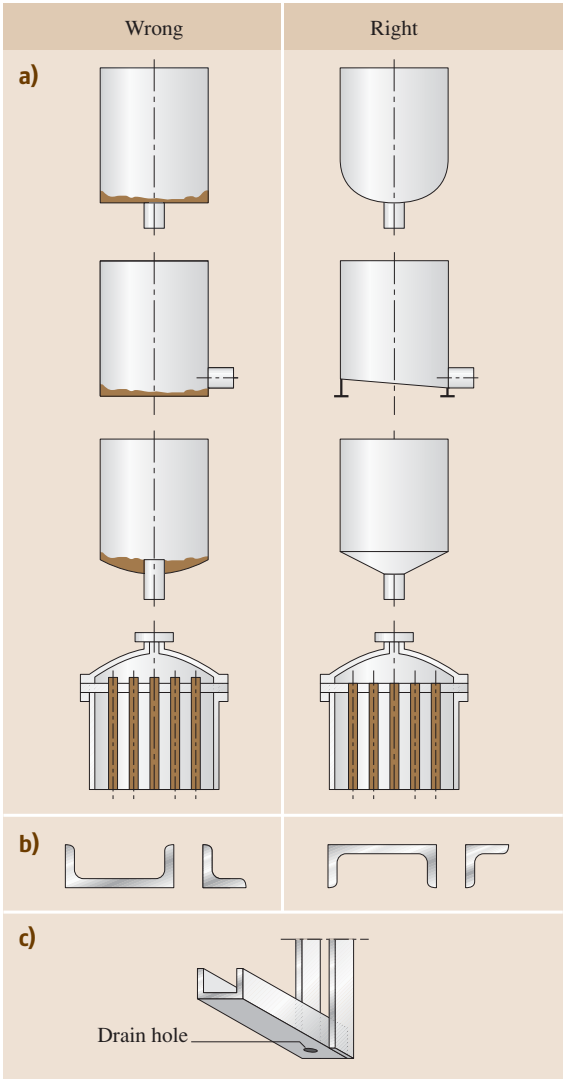


Fig. 9.32a–c Drainage of components susceptible to corrosion: **(a)** design of bases encouraging and impeding corrosion, **(b)** incorrect and correct arrangement of steel sections, **(c)** brackets made of channel section with drainhole (after [9.3])

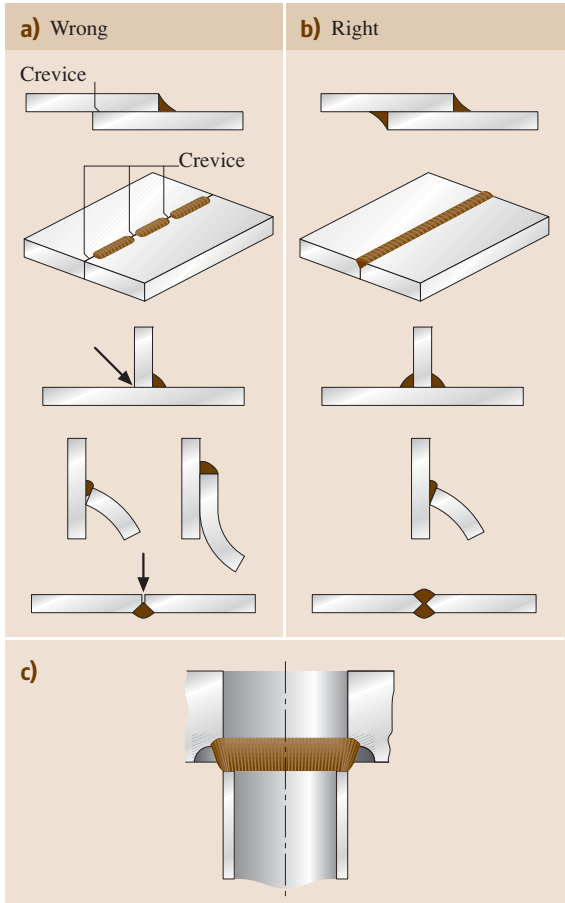


Fig. 9.33a–c Examples of welded joints: **(a)** susceptible to crevice corrosion, **(b)** correct design, **(c)** crevice-free welding of pipes, also improves resistance to stress corrosion cracking (after [9.3])

clamping or other compelling conditions exist, which would reduce the bearing strength of the structures. Channels are to be arranged in the direction of expansion or in the line of symmetry of the thermally or mechanically-conditioned distortion state of the component.

In the case of transient temperature changes, the thermal time constants of adjacent components are to be adapted in order to avoid to relative movements between these (components) [9.3].

Design in accordance with creep means taking into account the time-related plastic deformation of individual materials, especially at higher temperatures or the deformation of synthetic materials through the selection of materials and design, e.g. as much as possible avoiding a reduction in tension (relaxation) in stressed systems (screw connections, compressed connections) using elastic flexibility reserves. The field of tertiary creep is to be avoided through extremes of load and temperature, the selection of materials and stress time testing [9.3].

Design in accordance with corrosion includes avoiding the causes and/or preconditions for the vari-

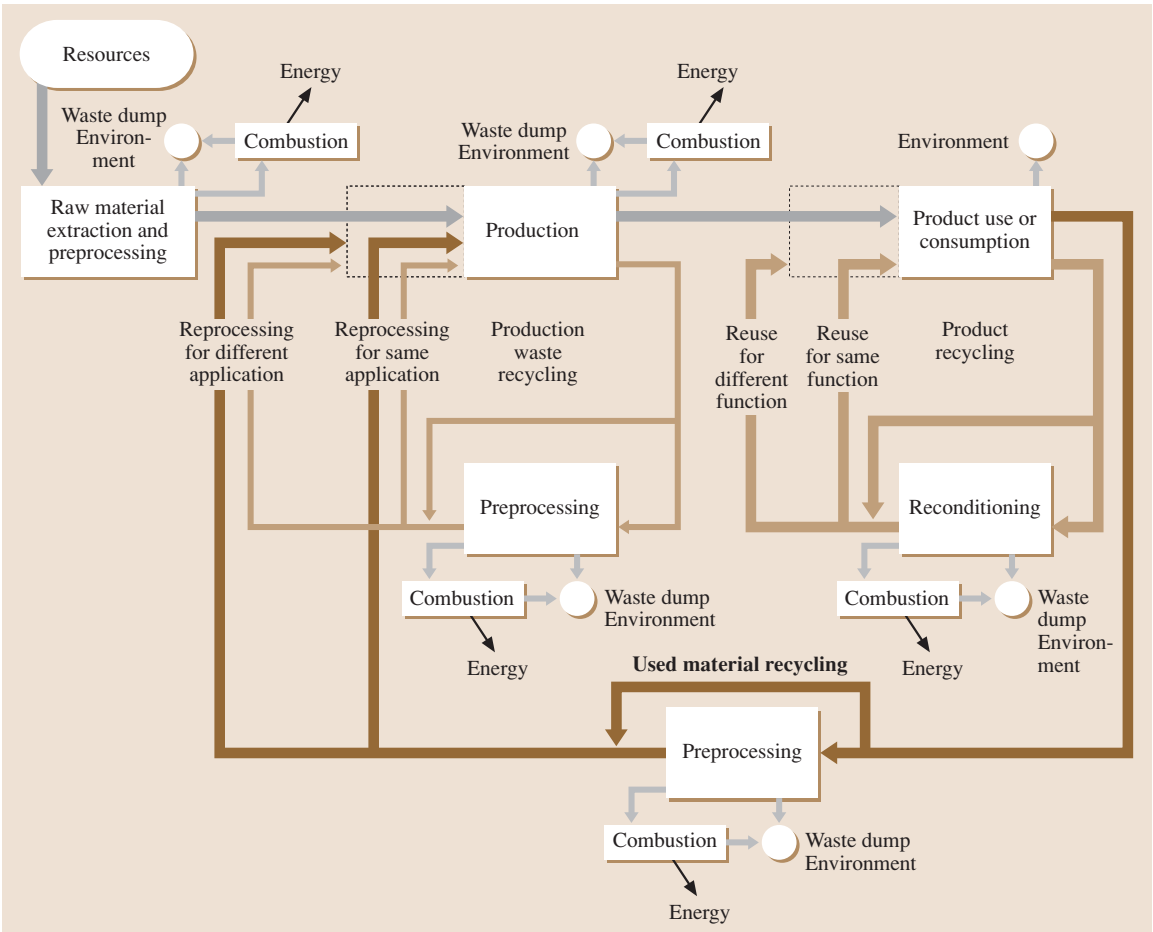


Fig. 9.34 Recycling options (after [9.3])

ous types of corrosion (primary measures) or through the selection of material, coatings or other protective/maintenance measures (secondary measures) which restricts the advancement of corrosion within permissible limits. Figure 9.32 illustrates constructive for avoiding moisture collection points and Fig. 9.33, examples of crevice corrosion.

Design in accordance with wear and tear means adopting the relative movement between parts necessary for the operation with as little wear and tear as possible through tribological measures in the system – material, surface or lubricant. At the same time, composite structures with high-strength boundary layers and design-giving basic materials provide an economic solution [9.32].

Design in accordance with ergonomic considerations includes taking into consideration the key characteristics, abilities and needs of the people to will use the product. In this, biomechanical, physiological and psychological aspects play a role. It is possible to make a differentiation between an active contribution of the person (e.g. in operating the product) and a passive relationship (retroactive and side effects as a consequence of the product) [9.33].

Design in accordance with shape (Industrial Design [9.23, 34]) means to take into consideration the fact that particular utensils not only serve a mere purpose but should also have aesthetic appeal. This is especially true for the look (shape, colour and labelling).

Design in accordance with manufacturing considerations means recognition of the significant influence of construction-related decisions on production costs, production times and manufacturing qualities and to take these into account in the optimizing components [9.3].

In designing parts (pieces) which are convenient to manufacturing processes, the designer has to be aware of the nature of the manufacturing procedures and the specific circumstances of each manufacturing plant (internal or external).

Design in accordance with assembly considerations means to reduce, to simplify, to unify and to automate the necessary assembly operations through an appropriate structure, as well as the design of the joints and joining parts [9.3]. In the design measures to simplify the parts production and the assembly, aspects of the testing process and production monitoring are looked at.

Design in accordance with norms include the norms which are observed for safety, usage and economic reasons and other technical rules which, as recognized engineering rules, serve the interests of manufacturers and users.

Design in accordance with transportation and packing considerations means taking into account standardized packaging and loading units (containers, pallets) for serial production as well as transportation options for large machinery [9.3].

Design in accordance with recycling considerations means knowing the nature of processing and reclamation procedures and supporting their use through assemblies and component design (shape, joints, materials). At the same time, reclamation-friendly constructive measures (facilitated dismantling and reassembling, cleaning, testing and post-processing or exchange) serve the interests of maintenance compatible design (inspection, servicing, repairs). Figure 9.34 shows recycling possibilities for material products, to which constructive measures must be oriented in order to facilitate recycling [9.35–38].

9.2 Basics

The methodical approach to the development and design of technical systems (engineering design) has established itself in virtually all design departments. Teaching specialized knowledge about methodical design is also a fixed component of the curriculum in the teaching of engineering sciences in universities and technical colleges.

There are a large number of approaches to design methodology, which are documented the technical literature. For example, *Ehrlenspiel* [9.1] focuses more on the cost approach to product development. One way of reducing and identifying costs early, according to Ehrlenspiel, is integrated product development. In his method on the other hand, *Roth* [9.2] divides the design process into many smaller steps and places strong emphasis on the incorporation of design catalogues in the solution process. *Pahl* et al. [9.3] worked very actively on the German guidelines VDI2221 [9.21] and VDI2222 [9.39] and subdivided the design process into individual activities, to which detailed methods are assigned. Further methods exist for these purposes, for example from *Koller* [9.6], *Gierse* [9.40], *Hubka* [9.41], *Bock* [9.42] and *Rugenstein* [9.43]. The essential aspect of each of these is the structuring of the task. This takes place, e.g., by drawing up flow diagrams and using methodical structuring aids, e.g., functional structures, efficacy structures or classification diagrams [9.44].

The methodical approach to the development of a technical system is clarified in this chapter using a practical example from the interdisciplinary field of biomedical engineering, based on the methodical method of *Pahl* et al. [9.3].

According to Pahl et al., the design process is divided into four stages:

- Precisely defining the task (problem identification)
- The concept stage
- The design
- Drawing up the final solution (detailed design)

As the example involves an interdisciplinary development project, it is particularly important to draw up only a few, but at the same time all, of the problem or work-related (sub)functions required for adequate structuring of the task and to represent these in a functional structure. It is also necessary to use a generally understood vocabulary. This enables us to ensure that people not yet involved in the process or people who do not have engineering training, e.g., medical experts or biologists can easily obtain an overview. This integration of employees from the individual specialized fields is necessary in order to be able to implement all medical and biological requirements at a high level.

9.3 Precisely Defining the Task

9.3.1 Task

The engineering system to be developed is a test setup for experiments with live human cells. The task (problem) for the designers was drawn up by the responsible medical experts. An extract from this is shown in the following.

For decades it has been known that certain cells in the human immune system are practically incapable of functioning in weightlessness. This can pose a serious problem for long-term stays in space on the International Space Station (ISS), or flights to Mars. The basic mechanism is to be investigated by means of weightlessness experiments with the help of parabolic (ballistic) flights. To this end, experimental equipment is to be designed with which tests on live cells can be performed onboard parabolic flights and in weightlessness. These experiments should also answer the question of whether humans are at all capable of living in weightlessness for any lengthy period. The findings can also be used in therapy for diseases of the immune system. It is necessary to mix the living human cells with an activator liquid and with a stopping liquid after a certain time. All the necessary safety requirements must be observed.

The designer's task consists of precisely defining this problem. This means that they must first draw up a functional engineering description. The aim is to draw up the whole function and all input and output variables for the engineering system to be developed.

9.3.2 Functional Description

The functional engineering description is drawn up by the responsible designer. It is used to clearly define the task or problem the designer has been set. At the same time it provides a basis for discussion with the other team members. In this way it is possible to identify early on whether there are any communication problems. In interdisciplinary projects it is particularly important to integrate the information of the non-engineering science team members into the technical preparations and therefore to create a basis for a methodical approach. It this project it was particularly important for the medical experts/biologists and engineers to speak the *same language*. The functional description is usually verbal. Frequently diagrams or initial sketches are also produced to depict the whole function to be fulfilled transparently. Figure 9.35 shows the outline technology for the test setup to be developed.

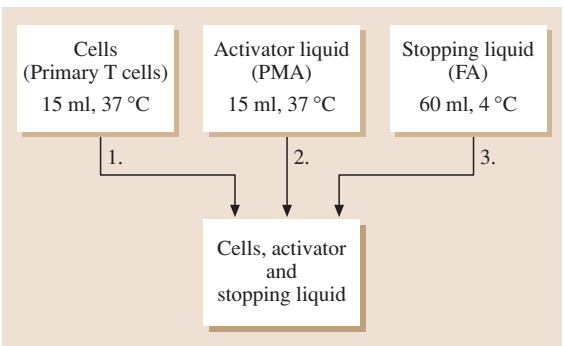


Fig. 9.35 Liquids to be mixed

This rough structuring was based on notes taken during team meetings and a functional structure drawn up by one of the medical–biological team members (Fig. 9.36).

This is already very finely structured. However, it is not drawn up in the usual form used in design methodology [9.3]. Further, such a precise description of a focused possible solution excludes other approaches and solutions in advance. The functional engineering description or the overall function to be fulfilled by the test setup can be described as follows.

A test setup is to be developed that enables three different cell lines to be mixed, to a large extent homogeneously, with certain activator liquids at the start of the weightlessness phase. Just before the end of the weightlessness phase a stopping liquid is to be added to the cell vessels filled with a cell type and an activator liquid.

In order to fulfil the specified medical requirements, combinations of three different cell liquids, three different activator liquids and two stopping liquids (Fig. 9.35) must be realized.

The condition of weightlessness was achieved with the help of parabolic flights. This means that an aircraft flies in a precisely defined parabola and the condition of weightlessness (micro-gravitation) is available for approximately 22–25 s (Fig. 9.37).

A main requirement is the fulfilment of all safety requirements in the test setup. Primarily, that under no circumstances may liquids escape from the test setup during the parabolic flights. Some of the cell lines used are genetically modified tumor cells and immune cells isolated from blood donators, and toxic liquids such as formaldehyde. These could pose a risk to the flight personnel during the weightlessness phase. This means

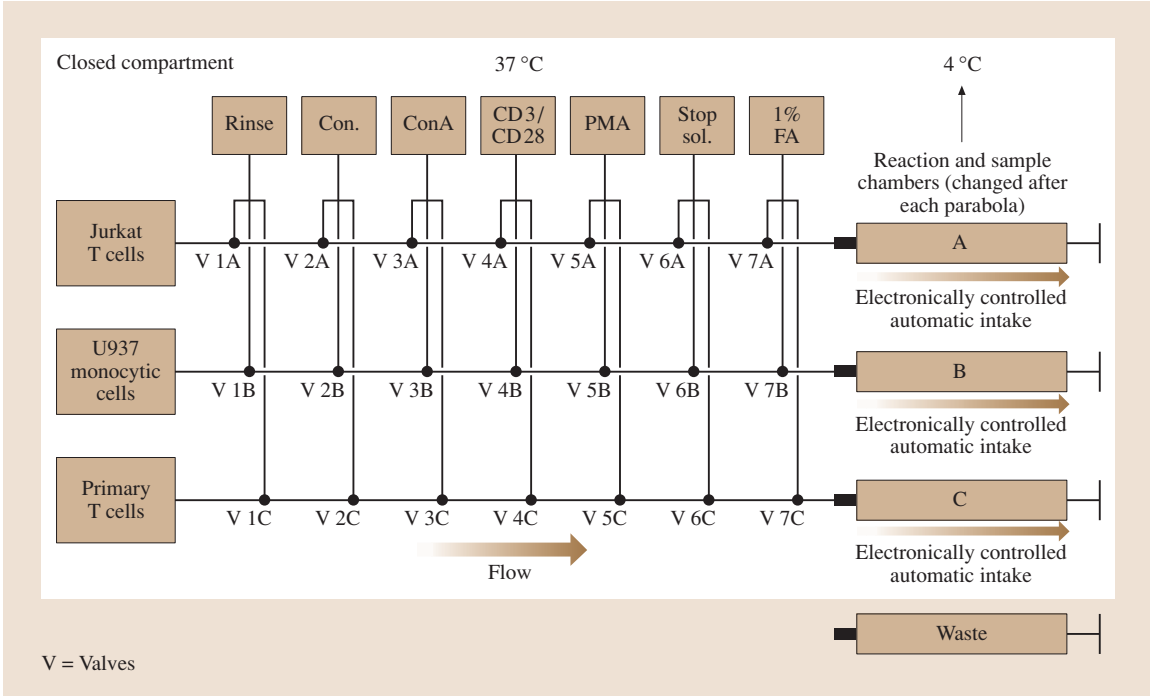


Fig. 9.36 Functional description from a medical point of view

that all parts that come into contact with the media or cell, activator or stopping liquids must be designed with double walls.

A further requirement is that the temperature of the cell and activator liquids must be 37 °C and the temperature of the stopping liquids must be 4 °C (Fig. 9.35).

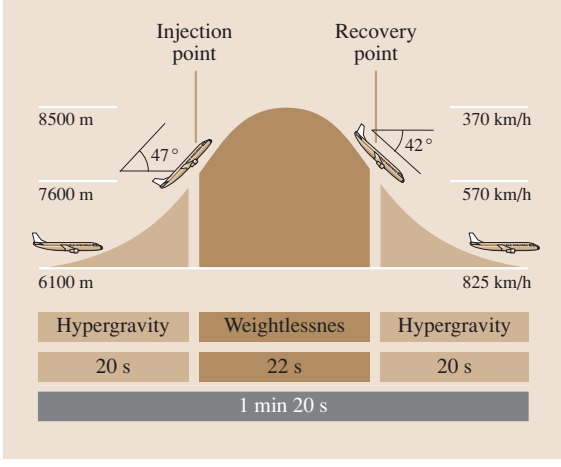


Fig. 9.37 Flight parabola for generating weightlessness (microgravitation) (after [9.45])

Further points included in an initial functional engineering description are:

- Enable fast and easy equipping with liquids
- Realization of the direct safety stage [9.3], i. e., leakproof under the conditions in the aircraft
- Clear functional sequences
- Good miscibility of the liquids during the experiment in the cell culture bag
- Fill under exclusion of air
- To a large extent transparent construction for observation of whether air inclusions exist
- Low weight (mass)
- Small space requirement
- Good cost effectiveness

This initial functional description is the basis for drawing up a requirements list.

9.3.3 Requirements List

When the task or problem is more precisely defined, other individual characteristic values and special requirements are determined. It is necessary to adequately describe all of the requirements set, both qualitatively and quantitatively. In this project this is achieved:

	Product: *Parabolic Flight*		Date: 06.02.06	Sheet 03
	Requirements			Source Responsibility
	No.	Descriptive information	Numerical info/comments	
Overall space required/ connection dimensions/ installation conditions		Aircraft door width	– 1.07 m	
		Aircraft door height	– 1.93 m	
		Cabin length	– 20 m	
		Maximum rack height	– 1,500 mm	
		Fixing points for experimental setup	– Mean rail spacing (y-axis) a) 503 mm b) 1006 mm – Hole diameter for screw M10 = 12 mm – Hole spacing in x direction = n * 25.4 mm > 20 inches (1 inch = 25.4 mm)	
		Maximum load per unit area over 1 m fixing rail length	– 100 kg	
		Rack structure	– Baseplate or frame connected to the seat rail system of the aircraft – There must not be any parts protruding from the baseplate in the direction of the flooring	

Fig. 9.38 Extract from the requirements list

- Through discussions with the other team members (biologists, medical experts)
- Through literature and patent research
- Analysis and evaluation of all applicable rules, regulations, etc. (technical requirements of the aircraft operator) [9.46]

The results of the precise task definition stage are documented in the requirements list. This usually contains the objectives to be realized and the prevailing conditions in the form of requirements and wishes [9.3]. The requirements must always be fulfilled. The wishes listed are to be realized if possible. The boundary between requirements and wishes can often not be clearly defined, especially in interdisciplinary projects. For this reason, such a differentiation was dispensed with for this project. An extract from the requirements list drawn up is shown in Fig. 9.38. At the same time, the requirements list is the legal basis for all further activities, including in this project.

9.4 Conceptual Design

In the conceptual design stage the overall function is structured. The result is a functional structure (Fig. 9.39). This means that the whole system is divided into its subfunctions and their links.

This procedure enables optimum analysis of the whole system. Efficacy principles were then assigned to the subfunctions.

Efficacy principles are usually based on physical effects that enable the function to be fulfilled. These are combined with geometric and material characteristics. In this project, conventional, intuitive and discursive solution-finding methods [9.3, 47] were used to draw suitable action principles. In detail:

- Conventional (e.g., literature or patent) research
- Intuitive (e.g., brainstorming)
- Discursive (e.g., the use of design catalogues)

When suitable efficacy principles have been determined for fulfilling the function, they are assigned to the subfunctions in a classification diagram. In this project the morphological box (Fig. 9.40) was used for this.

The efficacy principles drawn up to fulfil the individual subfunctions must then be purposefully linked to each other. When drawing up the concept for the test setup it was of primary importance that the high safety requirements be fulfilled with all the selected efficacy principles. This results in different efficacy structures. In practice it is usual to draw up a maximum of three efficacy structures. Figure 9.41 shows the path through the morphological box.

The efficacy structures generated are specified in greater detail and further developed to form basic solutions. The individual basic solutions are then assessed. An extract of the assessment (rating) undertaken in this project is shown in Fig. 9.42. The assessment criteria and assessment was carried out by the whole project team.

As a result, a basic solution was released to be drawn up. In general, as in this project, this is the ef-

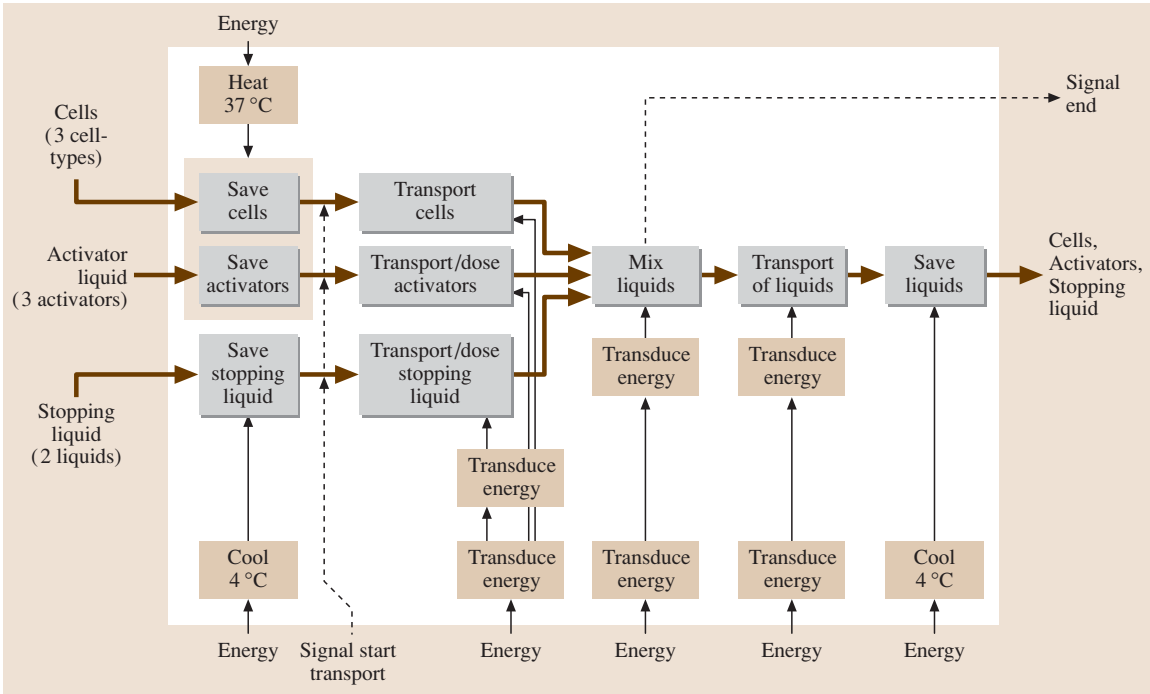


Fig. 9.39 Simplified functional structure

Function	Option	1.	2.	3.	4.
	Cool	Cooling accumulators Source: Novodirect	Peltier cooler Source: Rübsamen & Herr GmbH	Cryogenics	Refrigerator principle (compressor + heat exchanger)
	Heat	Sheet-type heating elements (silicon heating mats) Source: Hewid GmbH	Heat cartridges Source: Hewid GmbH	Infrared radiators Source: Hewid GmbH	Chemical reaction (thermal accumulators) Source: riedborn-apotheke
	Transport/ meter	Flexible tube pump Source: ismatec	Piston pump Source: Novodirect	Diaphragm pump Source: Novodirect	Gear pump Source: Novodirect
	Mix	Use of the pumps, pressure surge	Magnetic stirrer principle Source: Novodirect	Swivel movement of the vessels (shaker, vibrator)	

Fig. 9.40 Morphological box

ficacy structure with the best rating. It forms the basis of the design stage. This is shown in Fig. 9.43a.

The basic solution consists of two separate modules. The first module is the actual working module,

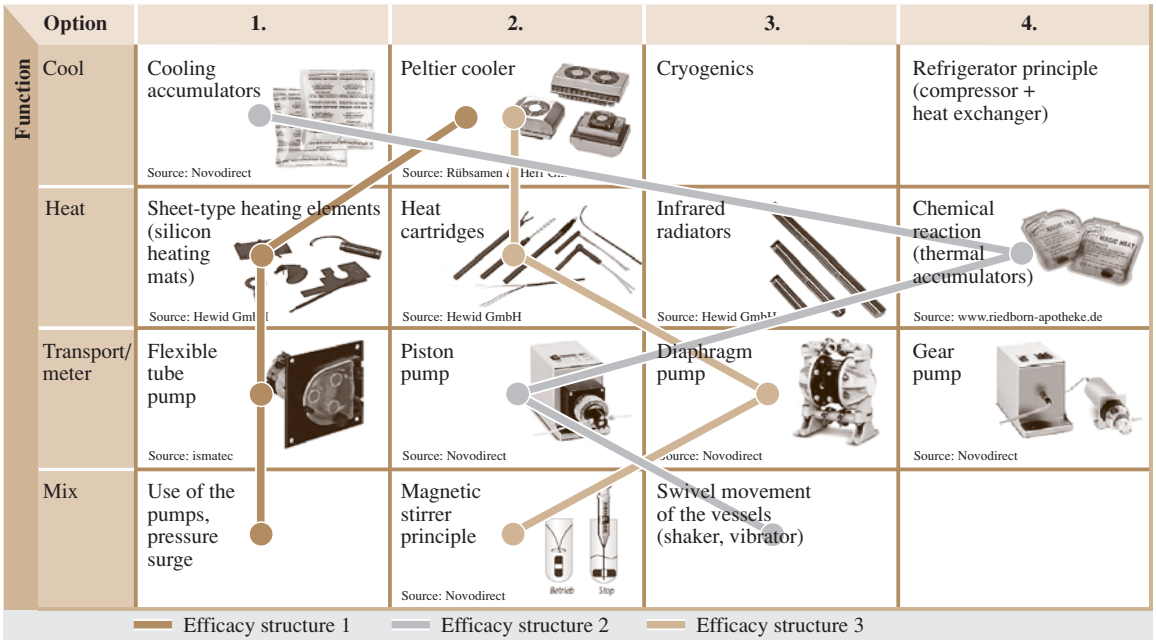


Fig. 9.41 Path through the morphological box

Assessment criteria	Weighting (W)	Opt. 1		Opt. 2		Opt. 3	
		Item (P)	W x P	Item (P)	W x P	Item (P)	W x P
37 °C uniformly distributed in the area of the cell storage and the activator liquids	0.8	4	**3.2**	1	**0.8**	3	**2.4**
4 °C uniformly distributed in the area of the stopping liquids and in the subsequent storage system	1.0	4	**4**	1	**1**	4	**4**
Low energy requirements	0.6	2	**1.2**	4	**2.4**	2	**1.2**
Low mass	0.7	3	**2.1**	3	**2.1**	2	**1.4**
Sterile pumping system with few mechanical components in area of contact with the pumped media	0.5	4	**2**	2	**1**	2	**1**
Total			30.0		25.3		27.1
Percent			0.83		0.70		0.75

Fig. 9.42 Extract from the assessment list

in which the cells, the activator and stopping liquids and all the necessary units are installed for pumping. This module is divided into three levels/submodules positioned on top of each other. Level 1 contains the pump for the stopping liquids and the cell vessels stored for filling, which is separated from the pump by a wall. Above this is the level for the power supply and controls. The top area contains the pump for the activators and, separated from this by a wall, the cell vessels to be filled are connected. After consulting the medical experts the information was received that three individual cell vessels are filled in parallel. The second module is the cooling module in which all filled cell vessels are stored at 4 °C after the experiment.

An important basis for this design is the joint specification from the medical experts and engineers in the project team that a previously precisely defined quantity of the cell liquid is already located in special cell vessels. The activator and stopping liquids are then pumped into these. The result is a simpler and better solution than the one previously proposed by the medical experts in Fig. 9.36. This arose as a direct consequence of the methodical approach described in Sect. 9.3.2. The new

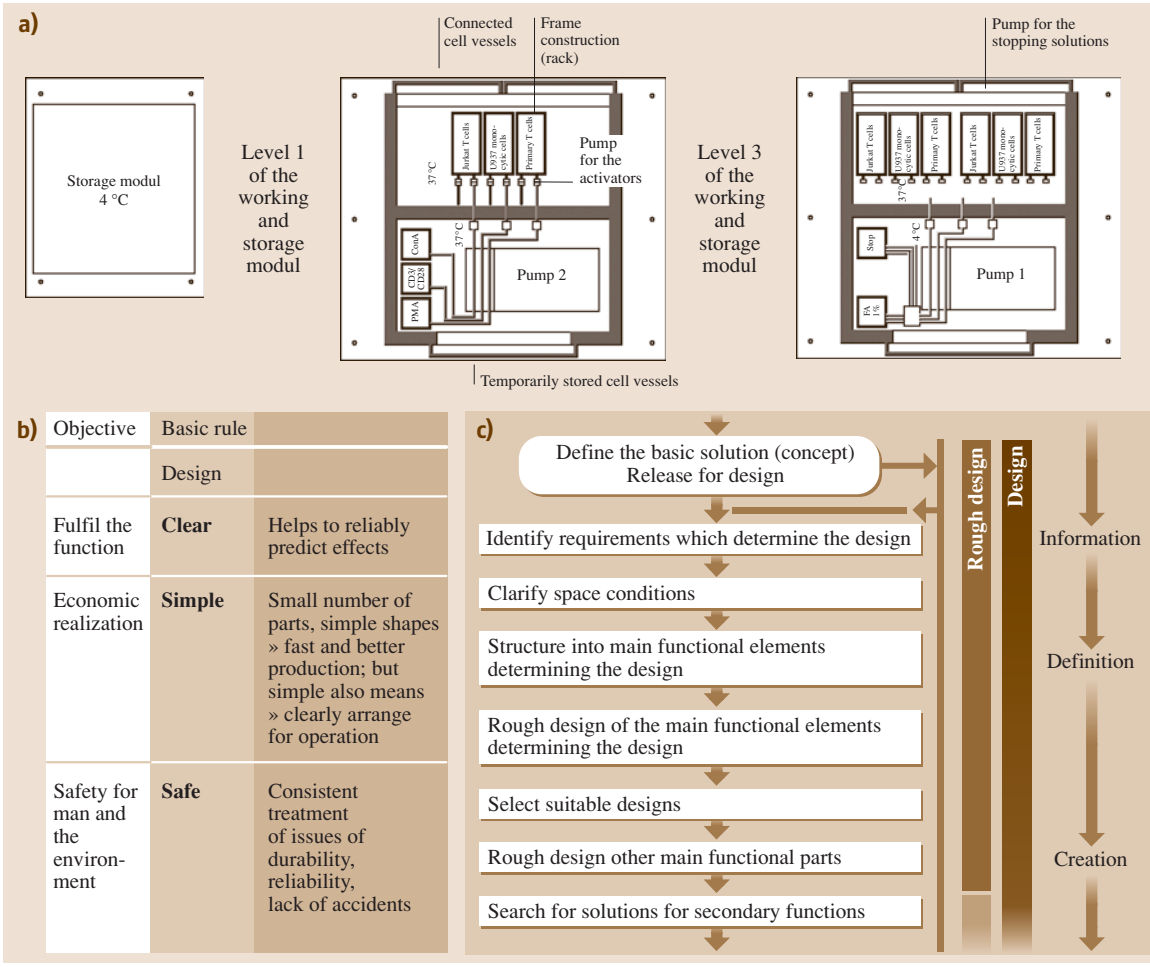


Fig. 9.43a–c Basic solution, released to be drawn up. (c) Extract from the main design activities (after [9.3])

solution prevents the cells themselves from being metered into the installed cell vessels by the pump, which would have generated shear forces that would have had a negative effect on the cells, exposing them to considerable stresses. In addition, it avoids repeated flushing of the pipes/lines for liquid transport. This fact thus minimizes the number of components (pumps, valves, lines) and therefore the costs incurred. In addition the costs for the liquids to be pumped are minimized (less flushing ⇒ less waste). This was an important aspect of *simply* fulfilling the basic design rule. The content is discussed in Sect. 9.5.

9.5 Design

The design stage is divided into:

- Rough design
- Detailed design
- Complete and check

The solution is more precisely defined during the design until a complete structure exists [9.3]. All the technical and economic requirements must be clearly and completely drawn up. The result is the design of the solution option, defining all the geometric, material

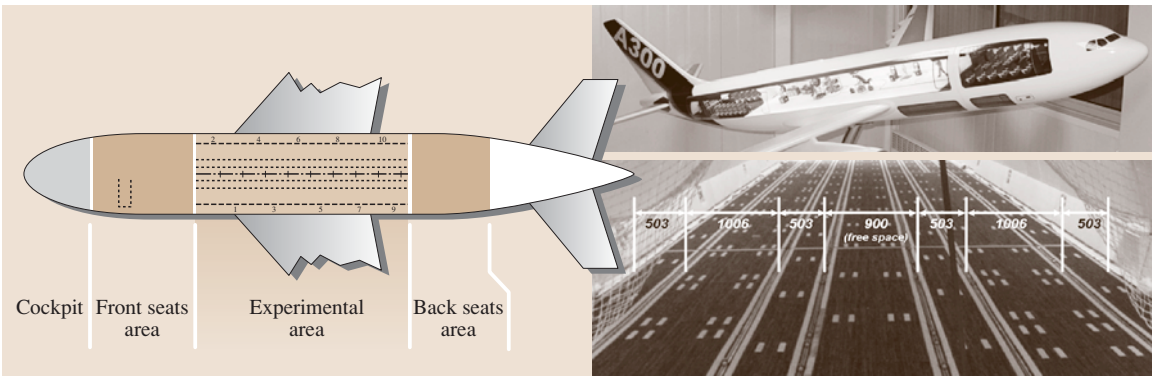


Fig. 9.44 Available (free) space and fixing options in the Airbus A300 of Novespace [9.45]

and condition characteristics. In this stage, the following three basic design rules must be observed: *simple*, *clear* and *safe* [9.3].

Figure 9.44 shows an extract of the main design activities.

The individual activities for developing the test setup for experiments with human cells are described in the following.

9.5.1 Identify Requirements that Determine the Design and Clarify the Spatial Conditions

The decisive requirements are essentially set by the ambient conditions, e.g., available space, effective and allowable stresses and loads, and the requirements set by the work sequence. The main requirements for the test setup to be developed are specified by the information contained in the aircraft operator's user manual. This document provided information on the internal dimensions of the aircraft frame and therefore the maximum effective heights and widths, the type and location of the fixing points, door dimensions for loading, the maximum allowable loads per unit area, details of the power supply, etc. (Fig. 9.44).

Requirements determined by the layout such as the flow directions and handling sequences were specified by the biomedical description of the experiment.

9.5.2 Structuring and Rough Design of the Main Functional Elements Determining the Design and Selection of Suitable Designs

In this activity a roughly structured diagram is drawn up for the main material flow. It names the preliminary main components selected. The main material flow is the pumping of activator and stopping liquids from storage into the cell vessel. Flexible tube pumps and suitable valves and hoses were selected for this task. The pump and valve sizes were chosen on the basis of the time and delivery rate requirements based on the biomedical process variables. Because of these specifications, instead of the originally planned flexible tube pump with a triple head for all activators and the same pump for all stopping liquids, six separate pumps had to be selected to achieve the objective.

Another main functional element was the frame (rack) for the modules. Extruded aluminium sections and accessories, available as a modular system and frequently used for automation engineering, were used in the design. The choice of section size depended on the calculated loads. Figure 9.46 shows the initial design of the working module.

9.5.3 Detailed Design of the Main and Secondary Functional Elements

The design of the main and secondary functional elements is a process that takes place in parallel in everyday design, as both groups may have a strong influence on each other. The pump–valve module (Fig. 9.49) is one of the main functional elements. Its decisive design requirements are those resulting from the biomedical process variables (size of the metered volume) and the boundary conditions resulting from the technical requirements (low mass, small space requirement, etc.)

A secondary functional element is the cell vessel which contains 15 ml of cell liquid at the beginning and into which the activator is injected before weightlessness starts, followed by the stopping solution after

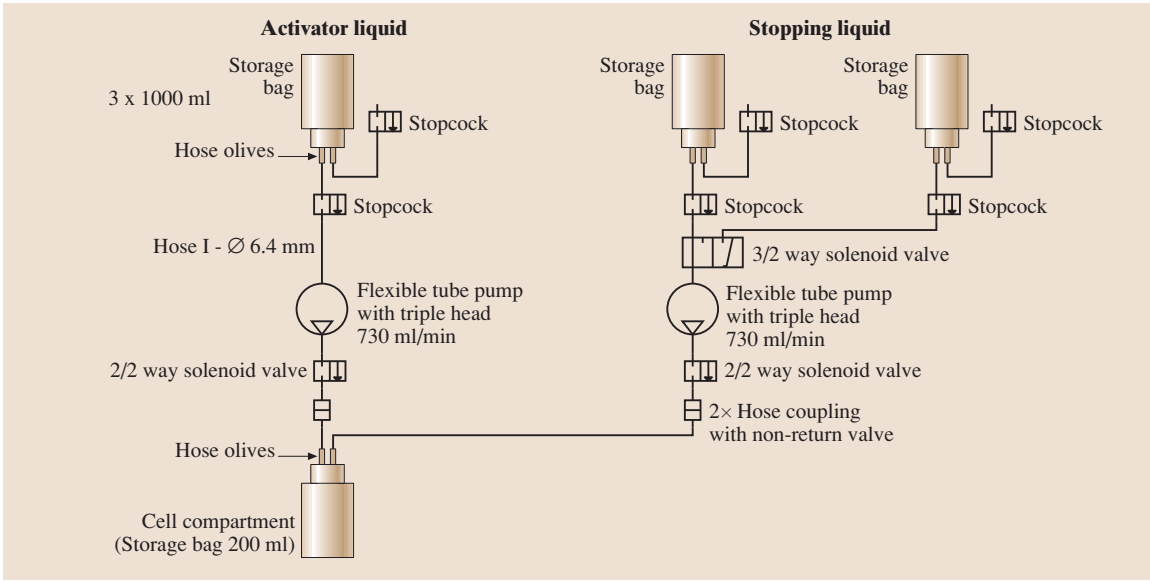


Fig. 9.45 Flow diagram for a cell vessel to be filled

approximately 22–25 seconds. The filling must be able to take place under the exclusion of air and in sterile conditions. Further, due to the safety requirements, this vessel must be designed with a double wall and enable fast removal of the contained liquids after the experiment. For biological and economic reasons the inner part of the vessel should be a one-off (disposable) product and the outer one should be reusable. Due to these requirements, further solutions were conceived and tested (Fig. 9.47).

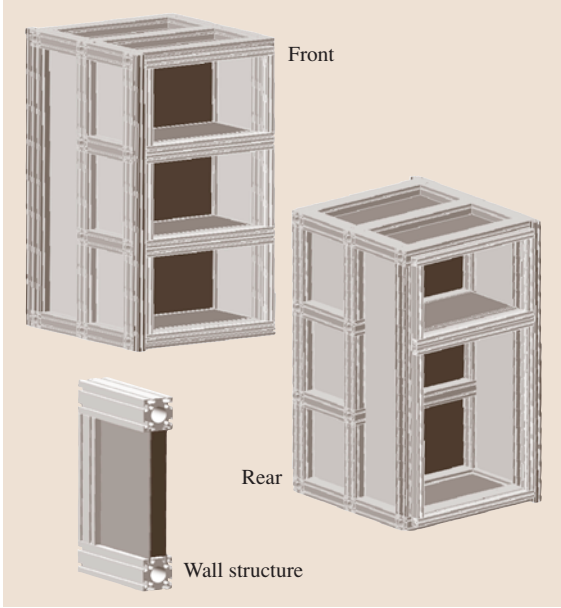


Fig. 9.46 Working module rack: front, rear, wall structure

Option 1 consists of an inner infusion bag integrated into a conventional 1 l plastic bottle. The connections are realized via the hose olives screwed into the bottle lid. Option 2 has a similar structure with a second liquid bag with a screw lid that provides the second wall. In the third solution the outer enclosure is formed by a specially produced plastic enclosure made using a rapid prototyping method.

The first two options have a very favorable price as all the components are production items. However they contain substantial defects in their functional fulfilment (filling under the exclusion of air). The reason for this is that, when the inner infusion bag is screwed in, it irreversibly twists. As a result, clear material flow is not possible, i. e., the basic design rule *clear* was not fulfilled. The third option is the most costly. However, it enables complete functional fulfilment according to the requirements. This is the preferred option and was released for design optimization. The result of the design using a continuous functional test during the optimization phase is shown in Fig. 9.48.

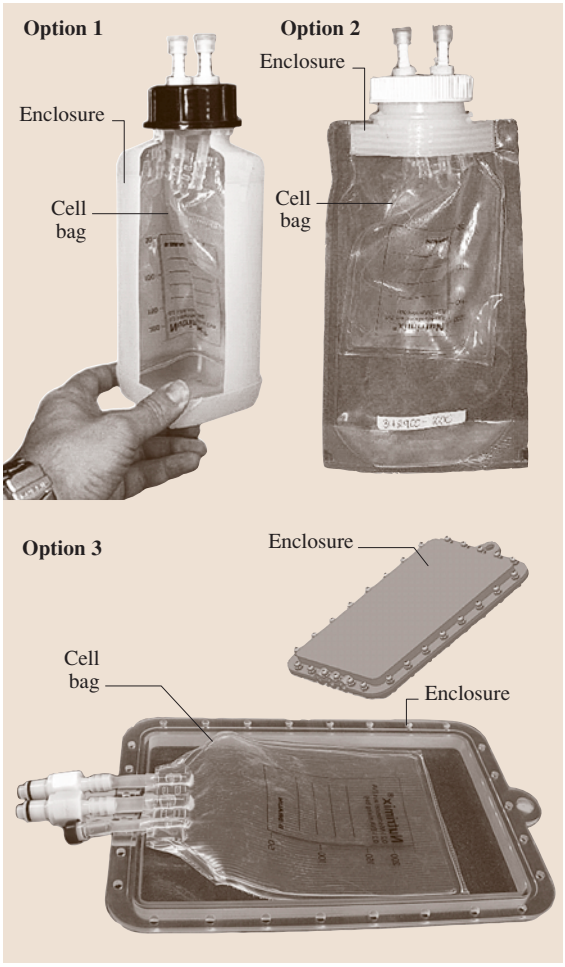


Fig. 9.47 Possible options for the secondary functional element: cell vessel (cell compartment)

9.5.4 Evaluation According to the Technical and Economic Criteria and Specification of the Preliminary Overall Design

During the design and associated continuously performed testing and control process it was found that individual technical requirements such as:

- Compliance with the maximum module dimensions
- Compliance with the maximum mass
- Compliance with the electricity consumption

could not be realized. Deviations from the requirements set in the requirements list were found.

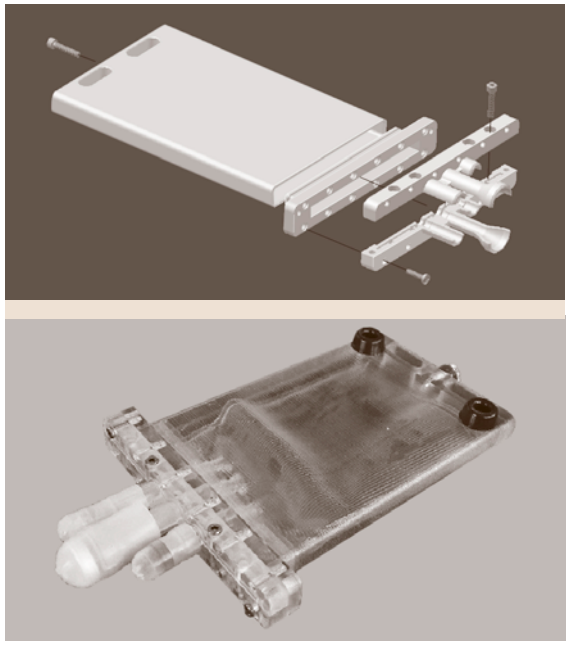


Fig. 9.48 Cell vessel structure

Furthermore, in this phase of the development work the functional fulfilment was checked. No deviations from the requirements list were found. The

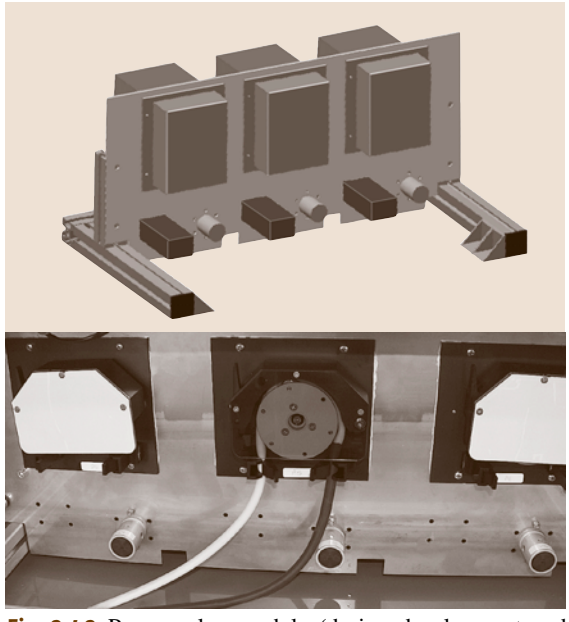


Fig. 9.49 Pump–valve module (during development and assembly)

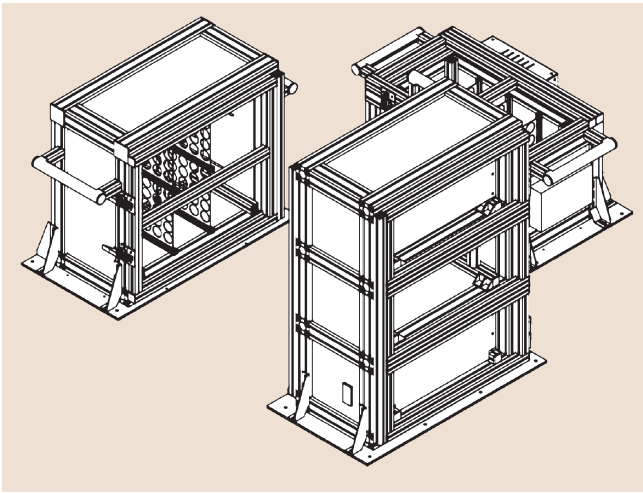


Fig. 9.50 Design for the experiment modules

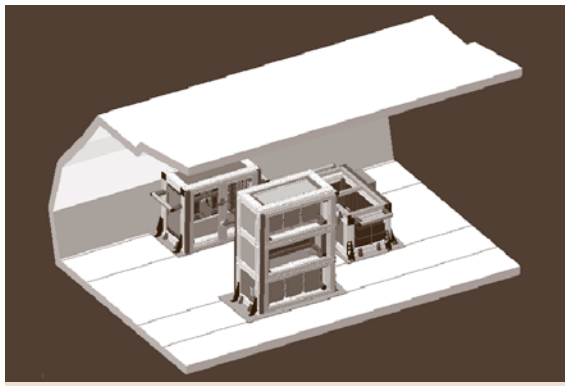


Fig. 9.51 Experiment modules

specified delivery rates of the pumps were fulfilled. The temperature ranges to be realized were achieved and the whole operational sequence was clear.

With respect to the economic criteria to be realized, there were also no deviations from the requirements list. All the specifications, such as material costs or production and assembly costs, were met.

A second design was drawn up based on the deviations from the requirements list. This design consisted of three separate modules (Fig. 9.50).

Module 1: The heating module for storing the cell compartments at 37 °C (incubator) before the experiment

Module 2: The actual working module, in which the cell vessels are filled

Module 3: The cooling module for storing the cell vessels after the experiment (4 °C)

This design was able to fulfil all of the technical and economic requirements and was released for further design work.

In the final phase of the design stage it is necessary to adapt the solution to existing standards and regulations. The individual components are assigned binding materials. During this phase, among other things, the drawings relevant for production are completed and the product documentation is produced. Figure 9.51 shows the result of the development.

9.5.5 Subsequent Consideration, Error Analysis, and Improvement

The main activities during the design phase according to [9.3] include the item: *checking for errors and disrupting effects*. This is a meaningful and necessary activity during design to prevent abortive development. However, systematic error analysis was only possible to a limited extent for the developed modules. Unlike other projects, in which empirical values already exist, process sequences are easy to follow or tests, or preliminary trials performed in parallel with the development process help to check for errors or faults, the analyses carried out for the experiment modules described here were to a large extent based on assumptions. During the development phase it was not possible to realize the condition of microgravitation for testing of the modules of the test setup. For this reason it was important to document and analyze the sequence and function of the modules during the parabolic flights. This was the only way to specifically enable error corrections and improvements. Several examples of modifications to the modules are listed in the following:

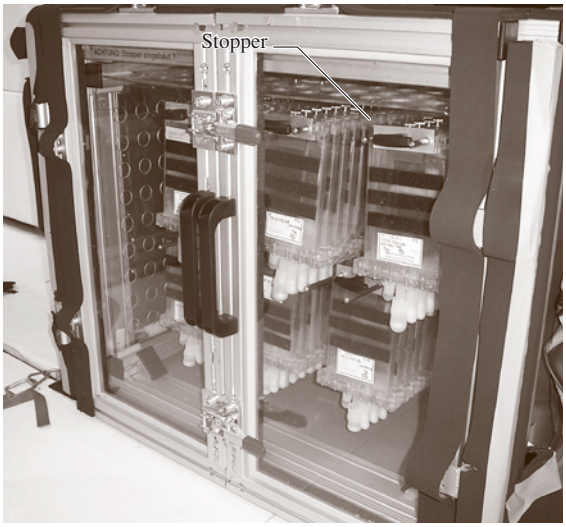


Fig. 9.52 Stopper in the heat module

- Most of the hoses from the medium replaced with rigid pipes
- Integration of safety sensors to identify the presence of the vessels to be filled before injection starts
- Replacing the manually opened venting valves in the cell vessels with automatically opening valves
- Improving the fixing (stoppers) of the cell vessels in the heating and cooling module

These modifications will be realized for subsequent flights.

The development and design of engineering systems according to methodical aspects based on information from the literature [9.1–3] is a useful procedure. The guidelines of design methodology were also applied in this interdisciplinary development project together with its tools, such as the requirements list, functional structure and morphological box, to name just a few. During the design phase of the product development process it was important to comply with the basic design rules: *simple*, *clear* and *safe* [9.3]. Several items that clearly show the realization of these three criteria are:

1. Simple:
   - Use of a module system for the rack design.
   - Only 15% of the required components are specially made (turned or milled parts).
2. Clear:
   - The liquid flow path is clear and does not lead to indefinable conditions.
3. Safe:
   - Redundant arrangement of parts absorbing forces and
   - Fixing of moving parts (cell vessels)

Primary importance was placed on realizing the direct safety requirements during the development activity. This task was successfully solved.

However, unlike the theoretical principles, practical experience shows that, especially during the conceptual and design phases, the experience and intuition of the designer are increasingly used to find the solution and systematic development is consciously dispensed with. This is not due to the fact that taught theoretical procedures are not generally practical, but to the increasing time and cost pressure for the development. It is often not possible for the designer to define several options for all main and secondary functions or to produce designs for the overall and part solutions and still produce a solution on schedule and within the cost framework. Here there is a risk that technically and economically better solutions are overlooked. One example from the project above is the fixing mechanisms (stoppers, Fig. 9.52) for limiting the sixth degree of movement for fixing the cell vessels in the working and heating module. Several optional solutions were not determined for this secondary functional element in advance, but instead the *first best* solution was used. In the technical test evaluations after the flights the operating personnel complained that due to the high stresses during parabolic flight these stoppers were difficult to undo and refasten. This solution had worked, but not optimally and will be changed for the next series of flights.

9.6 Design and Manufacturing for the Environment

The environment can be envisioned as interacting with human society in two ways: as a *source* of natural resources, and as a *sink* for emissions and wastes. The environmental problems addressed here are all related to overuse of both sources and sinks. Overuse of sources shows up as depletion and the reduced quantity and quality of resources. Overuse of sinks shows up as unbalancing of the harmony of previously natural pro-

Table 9.3 List of environmental concerns and links to manufacturing processes

Environmental concerns	Linkage to manufacturing processes
1. Global climate change	Greenhouse gas (GHG) emissions from direct and indirect energy use, landfill gases, etc.
2. Human organism damage	Emission of toxins, carcinogens, etc. including use of heavy metals, acids, solvents, coal burning
3. Water availability and quality	Water usage and discharges, e.g., cooling and cleaning use in particular
4. Depletion of fossil fuel resources	Electricity and direct fossil fuel usage, e.g., power and heating requirements, reducing agents
5. Loss of biodiversity	Land use, water usage, acid deposition, thermal pollution
6. Stratospheric ozone depletion	Emissions of chlorofluorocarbons (CFCs), hydrochlorofluorocarbon (HCFCs), halons, nitrous oxides, e.g., cooling requirements, refrigerants, cleaning methods, use of fluorine compounds
7. Land use patterns	Land appropriated for mining, growing of biomaterials, manufacturing, waste disposal
8. Depletion of non-fossil fuel resources	Materials usage and waste
9. Acid disposition	Sulfur and NO_x emissions from smelting and fossil fuels, acid leaching and cleaning

cesses. Often the change in balance takes years to detect and can be influenced by a variety of factors, making isolation and identification of the problems difficult and sometimes controversial. Nevertheless, over time many of these problems have been identified. They include ozone depletion, global warming, acidification, and eutrophication, among others. Corrective action often involves changes in the types and ways we use materials and energy for the production, use, and disposal of products. Table 9.3 lists commonly agreed environmental concerns and aspects of production, consumer use, and disposal that contribute to these concerns.

Table 9.3 clearly conveys the message that many of our environmental problems are directly related to materials usage, including energetic materials. In particular, note that several prominent concerns listed in Table 9.3 are directly related to our use of fossil fuels to generate energy. These include: CO_2 and NO_x emissions from the combustion of all fossil fuels, and SO_x and several heavy metals including As, Cd, Cr, and Hg, which are deposited onto land from the combustion of coal [9.48, 49]. In fact, at least four out of nine of the concerns listed above are related to fossil fuel use, including numbers 1, 2, 4, and 9. Because of this overriding importance, we will pay particular attention to tracking energy usage in the life cycle of products.

9.6.1 Life Cycle Format for Product Evaluation

A very important aspect of environmental analysis simply involves *connecting the dots*, in other words, showing the interconnectivity of human activities, and in particular, material flows. Few people contemplate where resources come from, or where they go after they are used, yet this is essential for life cycle analysis. With a life cycle accounting scheme one can then properly *burden* each product or activity with its environmental load. This information, in turn, can be used to answer the question, *is the utility gained from this product or activity worth the associated environmental load*? Although conceptually simple, this task is, in fact, quite complex. The major complexities are:

1. Establishing system boundaries
2. Obtaining accurate data
3. Representing the data with concise descriptors that appropriately assign responsibility
4. Properly valuing the results

Our approach will be to represent the product using material flow diagrams that capture the major inputs and outputs. In general, we will not attempt to relate these inputs and outputs to specific levels of environmental harm but only to identify them as *environmental loads*, known to cause harm, and which are excellent targets for technical improvement. When specific amounts of inputs used or outputs emitted are given, this type of analysis is called a life cycle inventory (LCI). The full life cycle analysis (LCA) includes LCI plus a connection between the loads produced and associated harm caused and often a ranking value among the different types of harm. Some LCA methods use these ranking values to generate a single number result. This can greatly ease decision-making, but requires agreement

with all of the implied value tradeoffs, something that is often difficult to accomplish.

Before proceeding further, it is important to more clearly establish the idea of a product life cycle. This is generally conceived as a materials flow process that starts with the extraction of raw materials from the Earth and ends with the disposal of the waste products back to the Earth. The general stages of this linear *once-through* cycle are:

1. Material extraction
2. Primary processing and refining
3. Manufacturing
4. Product distribution
5. Use
6. Final disposition

This sequence follows the principal product material flow, but of course there are multiple cross flows (consider the materials used by products, e.g., paper in printers and gasoline in automobiles) as well as back flows, such as product reuse, component remanufacturing, and material recycling. Figure 9.53 illustrates these flows in a general way, indicating cross flows both from nature and society as well as the major recycling flows. Society can then be represented by a vast array of these networks, interconnected but ultimately all originating from and leading to the ground – the Earth. This thought experiment clearly suggests the complexity of our problem. In practice this task is simplified by clearly defining the system boundaries and the objectives of the life cycle study. Problems can arise when the system considered is too large due to the interconnectivity of materials systems, and when the system considered is too small due to truncation. Matrix inversion methods, identical to those used in economic input–output analysis [9.50], along with high-level summary statistics have been called upon to help with the first problem [9.51, 52], while experience, iteration, and hybrid approaches are used to address the second [9.53, 54].

The commonest practice among LCA practitioners is based on developing process flow diagrams similar to Fig. 9.53 for the product, and tracing the major input and output paths to Earth. This requires data such as a bill of materials and lists of manufacturing processes, common use scenarios, distribution channels, and end-of-life characteristics for the product. The output is then a long list of material and energy inputs as well as emissions to the environment. These lists can easily include hundreds of materials, which then require some simplification and aggregation for interpretation. In this chapter, we will use a simplified format suggested by *Graedel* in his book on streamlined life cycle analysis (SLCA) [9.55]. This involves examining each stage of the life cycle and identifying major impacts and opportunities for improvement in five categories:

1. Materials choice
2. Energy use
3. Solid residues
4. Liquid residues
5. Gaseous residues

Graedel then suggests scoring each stage of the life cycle for each of the five categories with a numerical score from 0 (the worst) to 4 (the best). These scores are given relative to best practice for the product under consideration. In general, a score of 0 is reserved for a blatantly poor and/or uninformed practice that raises significant environmental concern, while a score of 4

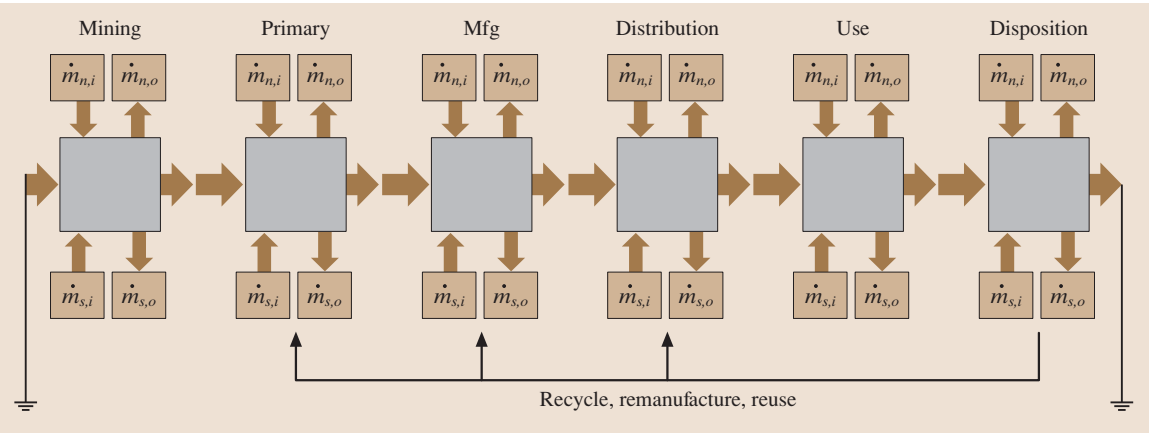


Fig. 9.53 Product life cycle material flows (m_n represent exchanges with nature, m_s represent exchanges with society)

Table 9.4 The environmentally responsible product assessment matrix [9.55]

Life cycle stage	Environmental stressor				
	Materials choice	Energy use	Solid residues	Liquid residues	Gaseous residues
Premanufacture	11	12	13	14	15
Product manufacture	21	22	23	24	25
Product delivery	31	32	33	34	35
Product use	41	42	43	44	45
Refurbishment, recycling, disposal	51	52	53	54	55

(The numbers are the indices for the matrix element M_{ij})

indicates excellent environmental performance with no known serious concerns. A perfect product would thus obtain a score of 100. Graedel gives more detailed guidance on how to score each element of the 5×5 matrix, as shown in Table 9.4, which represents the product.

9.6.2 Life Cycle Stages for a Product

In this section we will identify some of the major environmental issues that appear in each of the five stages of a product life cycle. The scoring of products for SLCA depends on the extent to which the designer and manufacturer make an effort to avoid these problems and substitute alternative materials and technology when possible.

Premanufacture: Materials Extraction and Primary Processing

Many of the environmental impacts associated with materials selection appear to occur in the very early stages of the material life cycle. This can be surmised by looking at United States national statistics for energy use, pollutants, and hazardous materials by various industrial sectors. For example, in Fig. 9.54, some of the manufacturing industries are broken down by standard industrial categories [standard industrial classification (SIC) codes] in terms of CO_2 and toxic materials per value of shipments. The primary processing of chemicals, petroleum and coal, and primary metals, have significantly larger environmental loads than other manufacturing sectors such as plastics and rubber, fabricated metals, machinery, electronics, and transportation. While not shown in Fig. 9.54, the metal mining industry would also show up prominently on this list. For example, toxic material releases for US metal mining in 1998 were equal to 145% of the toxic material releases from all of the manufacturing industries in the United States combined (including primary processing) [9.56].

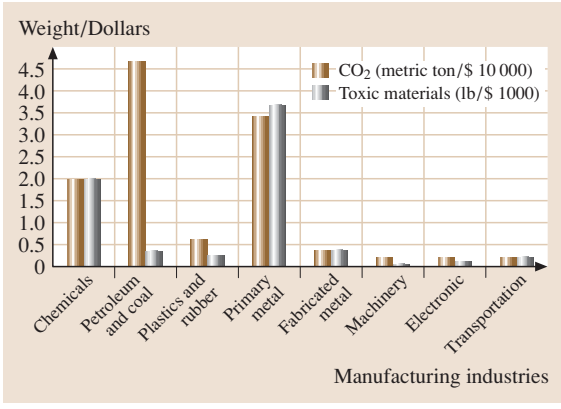


Fig. 9.54 CO_2 and toxic materials for several manufacturing industries

These large normalized impacts can be explained in two ways: relatively large emissions and relatively low prices. Primary processing industries handle very large quantities of materials, introducing many opportunities for economies of scale. At the same time, this high materials usage leads to high waste and emissions levels. For example, mining is very material intensive, producing ore waste-to-metals ratios that range from about 3:1 for iron and aluminum, to 10 000:1 for gold. In addition, many metals exist as, or occur in companion with, metallic sulfides. Once these materials are exposed to the surface they can oxidize into sulfates and sulfuric acid runoff, which can cause significant damage by acid mine drainage. Many of the commonest metals can lead to acid mine drainage, including copper, iron, nickel, lead, and zinc. In addition, some of the early processes can use other hazardous materials. If these materials escape, widespread environmental damage can occur. For example, the leaching of gold employs toxic cyanide compounds.

Table 9.5 Typical energy requirements for some common materials [9.57]

Material	Energy cost (MJ/kg)	Made or extracted from
Aluminum	227–342	Bauxite
Copper	60–125	Sulfide ore
Glass	18–35	Sand, etc.
Iron	20–25	Iron ore
Nickel	230–70	Ore concentrate
Paper	25–50	Standing timber
Polyethylene	87–115	Crude oil
Polystyrene	62–108	Crude oil
Polyvinylchloride	85–107	Crude oil
Silicon	230–235	Silica
Steel	20–50	Iron
Titanium	900–940	Ore concentrate
Wood	3–7	Standing timber

Table 9.6 Toxicity ratings for some of the elements [9.49]

Toxicity rating	Example elements
High toxicity	Beryllium, arsenic, cadmium, mercury, lead,
Moderate toxicity	Lithium, boron, chromium, cobalt, nickel, copper, bismuth
Low toxicity	Aluminum, silicon, titanium, iron, zinc, bromine, silver, tin, tungsten, gold,

Similarly, primary materials processing can be both materials and energy intensive. For example, the production of 1 kg of aluminum requires on the order of 12 kg of input materials and 290 MJ of energy [9.57]. The energy for this production plus other processing effects, in turn, leads to about 15 kg of CO_2 equivalent for every kg of aluminum produced [9.58]. Table 9.5 gives the energy requirements for some materials. Note that aluminum is in the high range of these materials, on the order of silicon but substantially less than titanium. The substitution of recycled materials can greatly reduce this energy requirement. Conversely the requirement for ultrahigh purity can greatly increase this requirement. For example, the recycled energy requirement versus *virgin* material is only about 5% for aluminum and 30% for steel [9.59], while the energy requirements for wafer-grade silicon used in the semiconductor industry is about 33 times that of commercial grade [9.60]. Hence, the mere act of selecting materials can in itself define a large part of the environmental footprint for a product. *Graedel* and *Allenby* suggest several other criteria to consider when selecting materials, including toxicity and abundance [9.49]. The ratings for some elements are given below in Tables 9.6 and 9.7.

Manufacturing Processes

As a group, manufacturing processes appear to be quite benign compared to materials extraction and primary processing, as indicated in Fig. 9.54. However, manufacturing processes often set many of the requirements for primary processing outputs. For example, processes with higher scrap rates require more energy in primary processing. Alternatively, processes that can use large quantities of recycled materials will have greatly reduced primary energy needs. This concept can be illustrated more rigorously by writing an equation for the embodied energy content for a hypothetical manufacturing process that uses E_m energy per kilogram of product produced. It has become common to discuss the energy "used up" in a process, but by the first law of thermodynamics we know that the energy is not actually lost. Rather, it is made unavailable. A more accurate thermodynamic quantity, *exergy* can be used up, and is more precisely what we mean in our discussion of *energy used*. Let the waste fractions be: α to ground, γ to recycle, and β to *prompt scrap* (recycled within the factory). This process uses a fraction ϕ of primary material with

Table 9.7 Classes of supply for some of the elements [9.49]

Worldwide supply	Example elements
Infinite supply	Bromine, calcium, chlorine, krypton, magnesium, silicon
Ample supply	Aluminum (gallium), carbon, iron, potassium, sulfur, titanium
Adequate supply	Lithium, phosphorus
Potentially limited supply	Cobalt[a], chromium[b], nickel[a], lead (arsenic, bismuth), platinum[b], zirconium
Potentially highly limited supply	Silver, gold, copper, mercury, tin, zinc (cadmium)

[a] Supply is adequate, but virtually all from South Africa and Zimbabwe. This geographical distribution makes supplies potentially subject to cartel control.
[b] Maintenance of supplies will require mining seafloor nodules. Note that materials in parentheses are co-mined with the parent material listed in front.

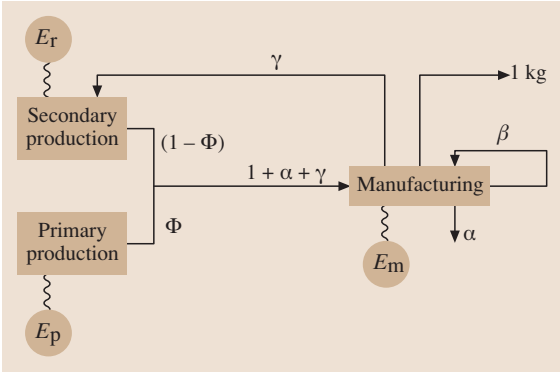


Fig. 9.55 System energy requirements for a manufacturing process

embodied energy E_p and a fraction $(1-\phi)$ of recycled material with embodied energy E_r, where in general $E_r \leq E_p$. From this, the sum of the energy requirement E_s, to produce one kilogram of product is (Fig. 9.55)

$$E_s = (\phi E_p + (1-\phi)E_r)(1+\alpha+\gamma) + E_m(1+\alpha+\beta+\gamma)\,. \quad (9.1)$$

Hence, even though it may be that $E_m \ll E_p$, (9.1) illustrates the long reach of the manufacturing process and its influence both up and down the product life cycle. As an example, consider the differences between machining and casting a part. While it might be true that $E_{m\,casting} > E_{m\,machining}$, generally speaking $\phi_{machining} \gg \phi_{casting}$. Furthermore, the waste for machining, particular α and γ, which show up in the first part of (9.1), can be quite large. In contrast, for large casting operations, most metal waste shows up in β, which does not occur in the first term. Hence in some situations, and quite counterintuitively, casting may be a more environmentally benign process than machining. Of course, this statement is based only on embodied energy usage and ignores other possible emissions.

Generally speaking however, while primary processing adds energy of order 100 MJ/kg (E_p) to any product, manufacturing adds energy of order 10 MJ/kg (E_m) [9.61,62]. The real role of manufacturing is that it draws in materials and energy not directly incorporated into the product and then expels them, often as wastes or emissions to the environment.

In addition to fossil fuel usage, a second environmentally important class of materials used in manufacture is cleaning fluids and coatings. Manufacturing often involves the cleaning and preparation of surfaces. Of particular concern are many of the solvents that are used to remove cutting fluids, lubricants, and other materials from the surface of the parts. In order to avoid the use of hazardous materials, many manufactures have replaced organics with water-based and mechanical cleaning methods [9.63].

Product Delivery

Product delivery involves two important types of environmental loads: transportation and packaging. The transportation of products around the world provides jobs and opportunities for many, but at the same time constitutes a major component of energy usage and related emissions. Furthermore, the geographical separation of product use from manufacturing can create significant barriers for the recycling of some materials.

Packaging waste is particularly egregious because of the large amounts of materials with only a very short intended lifetime. Furthermore, the customer gets the opportunity to experience this waste first hand.

Product Use

It is probably true that the vast majority of consumer appliances, electrical products, vehicles, lawn equipment, power tools, etc. – in short anything that has a power cord or runs on gasoline – has its largest or second largest impact during the use phase. Products with power cords draw energy from the utility station, which, in the US, have an average efficiency of about 35% and still burn 50% coal. These two facts alone often completely dominate the environmental impact of some products. Furthermore, powered devices can consume still other materials, e.g., paper and ink in printers, coffee in espresso machines, water in refrigerators with electric ice makers, etc. By and large, these automated appliances are considered desirable conveniences, but automated usage often, and unintentionally, leads to automated waste too.

Disposition

Most products in the USA end up in landfills, some are incinerated, and a few are recycled. In general, US landfill access has been significantly diminishing, particularly in the highly populated northeast. Many states have been closing landfills faster than they are opening new ones. Some states have moratoriums on new landfill development, and many ship their waste out of state. Furthermore, lined landfill sites for the collection of hazardous materials are highly restricted, leading to very high transportation and disposal costs for hazardous substances.

While incineration is not very popular in the USA, particularly in well-to-do communities, it is very much

an active option for a significant portion of the municipal solid waste (MSW) generated. Incineration can be combined with an electrical generation facility to produce power. Furthermore, the emissions can be scrubbed for various emissions. Nevertheless, it is difficult to tightly control the incoming waste stream and hence a wide variety of emissions, some not anticipated, can occur. In addition, it is well known that municipal incinerators are one of the top producers of dioxins in the USA, which are extremely expensive to scrub [9.64, 65]. Dioxins are a group of chemicals that have been found to be highly persistent, toxic, and carcinogenic.

A number of products are widely recycled in the USA. These include automobiles, tires (as a fuel to generate energy), newspapers, aluminum cans and, to a lesser extent, mixed paper and high-density polyethylene (HDPE) and polyethylene terephthalate (PET) bottles.

9.6.3 Product Examples: Automobiles and Computers

LCA, LCI, and SLCA can all help identify where major opportunities for environmental improvement occur for a product. The results depend on the product characteristics and the environmental loads of concern. Often our attention goes to those loads with the highest environmental profile. For example, in the life of a paper cup the use stage is short and the disposal can be benign, hence our attention goes immediately to the papermaking process, which has a variety of issues, many associated with the pulp bleaching process for kraft paper making, and possibly to the distribution stage, if the cups are transported a long distance. Another example, the disposable diaper, focuses attention on the waste disposal problem, while the reusable cloth diaper focuses attention on the energy-intensive washing cycle. Other products can be more complex, with major issues in several life cycle stages. Here we review life cycle issues for automobiles and computers.

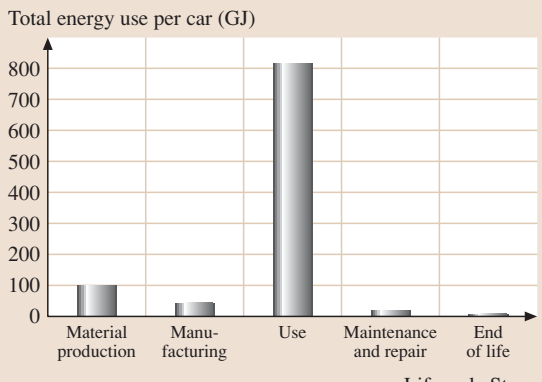


Fig. 9.56 Total energy use by life cycle stage for an automobile (after [9.66])

Table 9.8 Characteristics of generic automobiles [9.55]

Characteristics	ca. 1950s automobile	ca. 1990s automobile
Material (kg)		
Plastics	0	101
Aluminum	0	68
Copper	25	22
Lead	23	15
Zinc	25	10
Iron	220	207
Steels	1290	793
Glass	54	38
Rubber	85	61
Fluids	96	81
Other	83	38
Total weight (kg)	1901	1434
Fuel efficiency (miles/gallon)	15	27
Exhaust catalyst	No	Yes
Air conditioning	CFC-12	HFC-134a

Automobiles

The automobile has been subject to numerous studies concerning its environmental load [9.55, 66–71]. This discussion focuses on the automobile as a product. Other issues related to the automobile, for example how it has shaped our lifestyles and land use patterns, while very important, are not included in this discussion. As one can imagine, given the tens of thousands of parts, and the complexity of consumer behavior, vehicle types, and driving conditions, this is an enormous task. Yet, in spite of all of this complexity, the results have been quite consistent. By far the most important place to look for opportunities for environmental improvement is in the vehicle use stage. It is during this approximately 10 year period where the average vehicle, with a fuel efficiency of about 10 km/l (23.8 miles/gallon), burns about 14 metric tons of gasoline while traveling about 120 000 miles. Furthermore, due to the stoichiometry of the combustion process, this fuel consumption translates into some 40 metric tons of CO_2. When other aspects of the life cycle are included (the energy to make

Table 9.9 Premanufacturing ratings [9.55]

Element designation		Element value & explanation: 1950s automobile		Element value & explanation: 1990s automobile	
Materials choice	1, 1	2	Few hazardous materials are used, but most materials are virgin	3	Few hazardous materials are used, and much recycled material, Pb in battery in closed recycle loop
Energy use	1, 2	2	Virgin material shipping is energy intensive	3	Virgin material shipping is energy intensive
Solid residue	1, 3	3	Iron and copper ore mining generate substantial solid residues	3	Metal mining generates solid residues
Liquid residue	1, 4	3	Resource extraction generates moderate amounts of liquid residues	3	Resource extraction generates moderate amounts of liquid residues
Gas residue	1, 5	2	Ore smelting generates significant amounts of gaseous residues.	3	Ore processing generates moderate amounts of gaseous residues

Table 9.10 Product manufacture ratings [9.55]

Element designation		Element value & explanation: 1950s automobile		Element value & explanation: 1990s automobile	
Materials choice	2, 1	0	Chlorinated solvents, cyanide	3	Good materials choices, except for lead solder waste
Energy use	2, 2	1	Energy use during manufacture is high	3	Energy use during manufacture is fairly high
Solid residue	2, 3	2	Lots of metal scrap and packaging scrap produced	3	Some metal scrap and packaging scrap produced
Liquid residue	2, 4	2	Substantial liquid residues from cleaning and painting	3	Some liquid residues from cleaning and painting
Gas residue	2, 5	1	Volatile hydrocarbons emitted from paint shop	3	Small amounts of volatile hydrocarbons emitted

Table 9.11 Product delivery ratings [9.55]

Element designation		Element value & explanation: 1950s automobile		Element value & explanation: 1990s automobile	
Materials choice	3, 1	3	Sparse, recyclable materials used during packaging and shipping	3	Sparse, recyclable materials used during packaging and shipping
Energy use	3, 2	2	Over-the-road truck shipping is energy intensive	3	Long-distance land and sea shipping is energy intensive
Solid residue	3, 3	3	Small amounts of packaging during shipment could be further minimized	3	Small amounts of packaging during shipment could be further minimized
Liquid residue	3, 4	4	Negligible amounts of liquids are generated by packaging and shipping	4	Negligible amounts of liquids are generated by packaging and shipping
Gas residue	3, 5	2	Substantial fluxes of greenhouse gases are produced during shipment.	3	Moderate fluxes of greenhouse gases are produced during shipment

the fuel, etc.) and other greenhouse gases are converted into their CO_2 equivalent, the resulting equivalent CO_2 emissions over the life time of the vehicle are about 94 metric tons or 9.4 t/year. Other emissions during the use stage are also high, including NO_x, volatile organic compounds (VOCs), which contribute to ground-level ozone and smog, and other hazardous materials at lower levels. Other areas of concern are painting and cleaning during manufacturing, leaks and emissions during use and maintenance, and remaining quantities of unrecyclable materials: plastics, glass, foam, rubber, etc. The total energy use by stage, shown in Fig. 9.56, indicates that energy use during material production and manufacturing are also significant [9.66].

A general assessment of how the environmental performance of the automobile has changed over the years can be found in Graedel, who performed an SLCA to compare a 1950s automobile to one from the 1990s [9.55]. The assumed characteristics of the cars are given in Table 9.8. Their ratings for each of the five impact categories in each of the five life cycle stages are given in Tables 9.9–9.13. The final matrix values are summarized in Tables 9.14, 9.15, and plotted as a target plot in Fig. 9.57.

Computers

The study of the environmental footprint for computers is an interesting contrast to automobiles. While auto-

Table 9.12 Product use ratings [9.55]

Element designation		Element value & explanation: 1950s automobile		Element value & explanation: 1990s automobile	
Materials choice	4,1	1	Petroleum is a resource in limited supply	1	Petroleum is a resource in limited supply
Energy use	4,2	0	Fossil fuel energy use is very large	2	Fossil fuel energy use is large
Solid residue	4,3	1	Significant residues of tires, defective or obsolete parts	3	Modest residues of tires, defective or obsolete parts
Liquid residue	4,4	1	Fluid systems are very leaky	3	Fluid systems are somewhat dissipative
Gas residue	4,5	0	No exhaust gas scrubbing; high emissions	2	CO_2, lead (in some locales)

Table 9.13 Refurbishment/recycling/disposal ratings [9.55]

Element designation		Element value & explanation: 1950s automobile		Element value & explanation: 1990s automobile	
Materials choice	5, 1	3	Most materials used are recyclable	3	Most materials recyclable; plastics, glass, foam not recycled; sodium azide presents difficulty
Energy use	5, 2	2	Moderate energy use required to disassemble and recycle materials	2	Moderate energy use required to disassemble and recycle materials
Solid residue	5, 3	2	A number of components are difficult to recycle	3	Some components are difficult to recycle
Liquid residue	5, 4	3	Liquid residues from recycling are minimal	3	Liquid residues from recycling are minimal
Gas residue	5, 5	1	Recycling commonly involves open burning of residues	2	Recycling involves some open burning of residues

Table 9.14 Environmentally responsible product assessment for a generic 1950s automobile [9.55]

Life cycle stage	Environmental stressor					
	Materials choice	Energy use	Solid residues	Liquid residues	Gaseous residues	Total
Premanufacture	2	2	3	3	2	12/20
Product manufacture	0	1	2	2	1	6/20
Product delivery	3	2	3	4	2	14/20
Product use	1	0	1	1	0	3/20
Refurbishment, recycling, disposal	3	2	2	3	1	11/20
Total	9/20	7/20	11/20	13/20	6/20	46/100

Table 9.15 Environmentally responsible product assessment for a generic 1990s automobile [9.55]

Life cycle stage	Environmental stressor					
	Materials choice	Energy use	Solid residues	Liquid residues	Gaseous residues	Total
Premanufacture	3	3	3	3	3	15/20
Product manufacture	3	2	3	3	3	14/20
Product delivery	3	3	3	4	3	16/20
Product use	1	2	2	3	2	10/20
Refurbishment, recycling, disposal	3	2	3	3	2	13/20
Total	13/20	12/20	14/20	16/20	13/20	68/100

mobiles use mostly conventional materials and many standard manufacturing processes, the microchips in computers use much more-specialized materials and rapidly changing process technology. The result is that the complete life cycle of the computer has not been filled in to the extent that the automobile has. This is clearly illustrated in the important paper by *Williams*, *Ayres*, and *Heller* [9.60], which illustrated that there is far less agreement on the magnitudes of the environmental impacts associated with microchip fabrication.

Nevertheless the available data indicate that microelectronics fabrication is very materials and energy

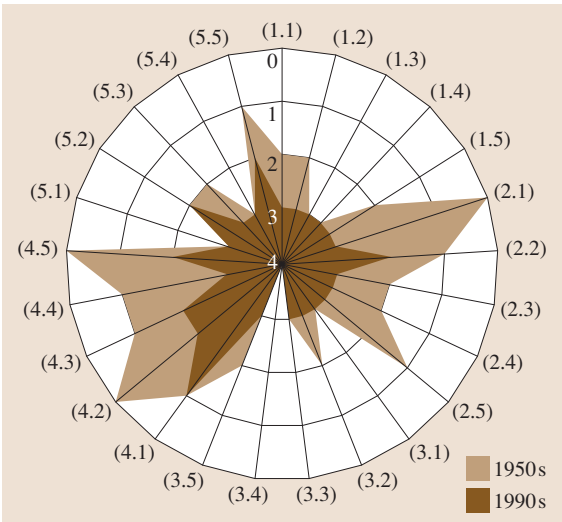


Fig. 9.57 Target plot of the environmental impact of generic automobiles for the 1950s and 1990s, see text (after [9.55])

intensive, in fact orders of magnitude more so than, for example, automobile manufacturing. In particular, approximately 1.7 kg of materials (including fuel) are needed to make a 2 g microchip. This certainly undermines some claims that microelectronics represent a form of dematerialization. Some of the chief findings of Williams et al. are summarized in Table 9.16.

At the same time, it is important to keep in mind that a computer is made up of much more than microchips, and the life cycle includes more than the fabrication stage. An approximate bill of materials for a desktop computer and cathode ray rube (CRT) monitor are given in Table 9.17. In this context the microchips and their constituents hardly show up. In fact many of the materials used in a computer are rather conventional. However, there are some materials of concern. Lead is present in both the central processing unit (CPU) and the monitor, cadmium (not listed) may be present in the batteries, mercury can be used in some switches

Table 9.16 Chief findings for microelectronics fabrication for a dynamic random-access memory (DRAM) chip [9.60]

Mass of 32 MB DRAM chip	2 g
Total chemical inputs	72 g/chip
Total fossil fuel inputs	1600 g/chip
Total water use	32 000 g/chip
Total elemental gas use	700 g/chip

Table 9.17 Bill of materials for a desktop computer and CRT monitor [9.72]

Material	Mass (g)	Use
Desktop computer		
Steel	6050	Housing
Copper	670	Wires, circuit boards
Aluminum	440	Housing, CD ROM
Plastics	650	Circuit boards
Epoxy	1040	Solder
Tin	47	Solder
Lead	27	Disk drive
Nickel	18	Circuit boards
Silver	1.4	Circuit boards
Gold	0.36	
Subtotal	8944	
Other	96	
Total	9040	
17" CRT monitor		
Glass	6817	Picture tube
Steel	2830	Housing
Copper	700	Wires, circuit boards
Ferrite	480	Deflection yoke
Aluminum	240	Heat sinks
Plastics	3530	Housing
Epoxy resin	140	Circuit boards
Tin	20	Solder (circuit boards)
Lead	593	Glass, solder
Silver	1.24	Circuit boards
Gold	0.31	Circuit boards
Subtotal	15352	
Other	98	
Total	15450	

and is used in laptop displays, and there is growing concern, particularly in Europe, over brominated flame retardants, which are used in the plastics.

Currently, there is active work to develop a complete life cycle analysis available for computers. An earlier study looked at energy, waste, hazardous materials and water used in the life cycle of a computer workstation [9.55,73], and more-recent information particularly on the fabrication, use, and end of life stages for a personal computer is given by *Kuehr* and *Williams* [9.72]. Computers seem to raise concerns at all stages of life. The premanufacture and manufacturing stages are very material and energy intensive due primarily to the high material purity needed for the microelectronics fabrication. Distribution can be a concern due to the large amount of packing materials and the long dis-

Table 9.18 Streamlined life cycle analysis: desktop computer display and CPU. Premanufacturing

i, *j*	Environmental stressor	Score
1, 1	Material choice Few recycled materials are used. Many toxic chemicals are used, including Pb in CRT and PWB, Cd in some batteries, Hg in some switches, and brominated flame retardants in plastics.	0
1, 2	Energy use Extra-high-grade materials for microchip very energy intensive. Other high-energy materials include virgin aluminum, copper, CRT glass.	1
1, 3	Solid residues Many materials are from virgin ores, creating substantial waste residues. Si wafer chain is only 9% efficient.	1
1, 4	Liquid residues Some metals from virgin ores can cause acid mine drainage.	2
1, 5	Gaseous residues Very high energy use and other materials use lead to substantial emissions of toxic, smog-producing, and greenhouse gases into the environment.	1

Table 9.19 Streamlined life cycle analysis: desktop computer display and CPU. Product manufacture

i, *j*	Environmental stressor	Score
2, 1	Material choice Manufacturing uses restricted and toxic materials (see 1,1) plus cleaning solvents.	1
2, 2	Energy use Energy use in production is very high for ICs and PWB and moderate to high for conventional materials. If we examine energy use during the manufacture of individual parts of the computer: microchip (0), printed circuit board (1), cathode ray tube (2), LCD (0), other bulk material (3)	1
2, 3	Solid residues There are large solid residues for chemical processes, such as CVD, PVP and plating, e.g., printed circuit boards yield 12 kg of waste for each kilogram of finished product. Also high performance requirements often result in low yields.	1
2, 4	Liquid residues Large quantities of waste liquid chemicals, e.g., approximately 500 kg of waste liquid chemicals for each kilogram of product including plating solutions and cleaning fluids. Very high volumes of water are also used.	1
2, 5	Gaseous residues Manufacturing energy use and processes lead to substantial emissions of toxic, smogproducing, and greenhouse gases into the environment.	1

Table 9.20 Streamlined life cycle analysis: desktop computer display and CPU. Product packaging and transport

i, *j*	Environmental stressor	Score
3, 1	Material choice Several materials, large quantities, minimal recycling activity.	3
3, 2	Energy use Long distances traveled. Large volumes of materials.	2
3, 3	Solid residues Waste volumes are large, no arrangements to take back product packaging after use.	2
3, 4	Liquid residues Little or no liquid residue is generated during packaging, transportation, or installation.	4
3, 5	Gaseous residues Gaseous emissions are released by transport vehicles.	2

tances some products need to travel. The use phase can be very energy intensive. For example, data given by *Kawamoto* [9.74] and *Cole* [9.75] sets the residential annual energy use for a desktop computer and monitor at about 380 MJ/year. In a commercial/industrial setting where the computer monitor may be on con-

Table 9.21 Streamlined life cycle analysis: desktop computer display and CPU. Product use

i, *j*	Environmental stressor	Score
4, 1	Material choice Power from electrical grid uses 50% coal.	2
4, 2	Energy use High to very high energy usage.	1
4, 3	Solid residues Little direct solid residues (excluding printing functions) but power uses coal, resulting in mining residues.	3
4, 4	Liquid residues Little direct liquid residues (excluding printing function) but coal mining yields liquid residues.	3
4, 5	Gaseous residues No emissions are directly associated with the use of computers. However, gaseous emissions are associated with energy production for use of computers.	1
Does not include printing		

Table 9.22 Streamlined life cycle analysis: desktop computer display and CPU. Refurbishment/recycling/disposal ratings

i, *j*	Environmental stressor	Score
5, 1	Material choice Product contains significant quantities of lead and brominated flame retardants and may contain mercury and cadmium. Often these are not clearly identifiable or easily removable. Many materials are not recycled.	1
5, 2	Energy use The product is not designed for energy efficiency in recycling, or for high-level reuse of materials. Also, the transport of recycling is energy intensive because of weight/volume and location of suitable facilities.	1
5, 3	Solid residues Dissimilar materials are joined together is ways that are difficult to reverse and the product overall is difficult to disassemble. Little recycling. Short life cycle of computers compounds these problems.	1
5, 4	Liquid residues Product contains no operating liquids and minimal cleaning agents are necessary for reconditioning (not including printing functions).	3
5, 5	Gaseous residues Roasting of printed wiring boards (PWBs) to recycle metals leads to emissions.	2

Table 9.23 Streamlined life cycle analysis: desktop computer display and CPU. Environmentally responsible product assessment for the computer display and CPU

Life stage	Materials	Energy	Solid	Liquid	Gaseous	Total
Premanufacture	0	1	1	2	1	**5/20**
Product manufacture	1	1	1	1	1	**5/20**
Product delivery	3	2	2	4	2	**13/20**
Product use	2	1	3	3	1	**10/20**
Recycling	1	1	1	3	2	**8/20**
Total	**7/20**	**6/20**	**8/20**	**13/20**	**7/20**	**41/100**

tinuously, the estimate is 1500 MJ/year. For a 3 year lifetime, this is more than half of the energy used in production. The end-of-life issues are several; there are several *materials of concern* as mentioned earlier, and the shear volume of retired computer and electronics represents a significant solid waste/recycling challenge. Currently only a very small percentage of computers are recycled [about 11%, as estimated by the Environmental Protection Agency (EPA)]. Using the materials lists given in Table 9.17 and the information cited in the references reviewed here [9.55, 60, 72–76] we have developed a baseline SCLA for a 1990s desktop computer and CRT display. The results are given in Tables 9.18–9.23. A target plot is given in Fig. 9.58.

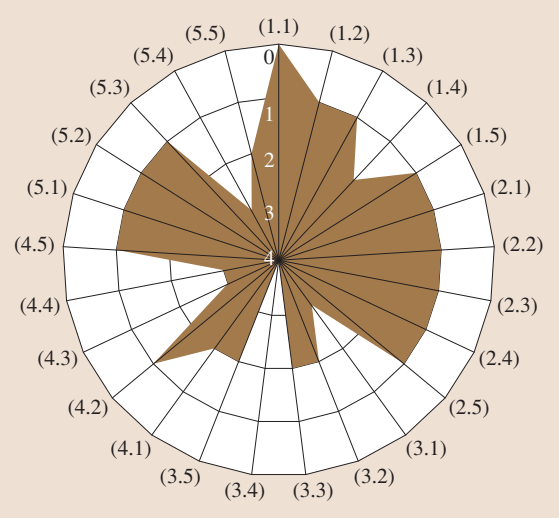


Fig. 9.58 Target plot of the environmental impacts of an early-1990s desktop computer and CRT display, see text (after [9.55])

Using Life Cycle Information for Product and Process Improvement

One of the primary reasons for developing life cycle environmental information is to identify opportunities for improvement. Streamlined life cycle analysis (SLCA) as developed in the two preceding sections is particularly good at this because it provides an overview with an emphasis on high-profile issues. The automobile example illustrates important environmental improvements from the 1950s to the 1990s, at the same time indicating several remaining challenges, particularly related to fossil fuel usage.

The SLCA given for the 1990s desktop computer is intended as a baseline for identifying critical areas for improvement, which are many. In comparison, the current laptop computer would score much more favorably. Important improvements would include:

1. A drastic reduction in the total weight of materials used
2. Elimination of significant amount of lead found in the CRT (although small amounts of mercury are needed for the laptop display)
3. Significantly reduced packing materials
4. Significantly lower energy required in usage

The SLCA methodology presented here is a compact way to learn about a complex problem, however, once a significant target for environmental improvement has been identified by SLCA, a more quantitative life cycle inventory for the material(s) of concern would be warranted. This would allow a much more precise evaluation of potential improvements; for example, Fig. 9.56 shows the total energy used for a generic family automobile in the late 1990s. An LCI study would then connect each of these energy requirements to the energy technology used and the corresponding emissions. For example while the use phase burns gasoline in an internal combustion engine, the material production and manufacturing phases use a mixture of mostly coal along with natural gas, oil, and even nuclear (from the electricity grid). Hence connecting the energy requirements to the fuel cycles would lead to specific contributions of each type of pollutant of interest. For example, energy from electricity in the USA comes from 50% coal, 20% nuclear, 18% natural gas, 7% hydroelectric, 3% oil, and 2% renewables. The corresponding emissions per electricity used are: 633 kg CO_2/MWh, 2.75 kg SO_2/MWh, 1.35 kg NO_x/MWh, and 12.3 gmercury/GWh, where MWh is a megawatt-hour, and GWh is a gigawatt-hour [9.77].

A full-blown life cycle analysis (LCA) would include weightings of the different pollutants based on human value judgments. For example, the LCA software Simapro from the Netherlands employs a weighting scheme called the *Eco-indicator* that reflects the value judgments of Europeans concerning the various environmental issues that concern them. Obviously such schemes could vary widely depending upon local concerns.

Before leaving this section it is worth noting that, because SLCA yields numerical scores for products, it might be tempting to compare different types of products. This is possible but must be done with caution. In the first place the SLCA methodology developed by Graedel and Allenby was intended as a relative indicator for a particular product type. Therefore the methodology would have to be modified. A second problem is that dissimilar products can have vastly different utilities for the users. Hence the comparison may seem nonsensical to some. Nevertheless some comparisons can be useful. For example, Williams has pointed out that the fossil fuel used during the entire life cycle of the production and use for a computer is almost identical to that of a refrigerator even though the computer only lasts about one quarter of the time (2.5 years versus 10 years) and the refrigerator is on all of the time [9.76]. The obvious implication is that much more could be done to improve the energy efficiency of computers.

Table 9.24 Energy efficiency guidelines [9.49, 78, 79]

Action	Reason
Do SLCA/LCI/LCA for product	Identify energy usage
Encourage use of clean renewable energy sources	Reduce harmful by-products and preserve resources
Choose the least harmful source or energy Note for fossil fuels the cleanest is natural gas followed by oil products, and then coal	Reduce harmful by-products
Have subsystems power down when not in use	Reduce energy usage and fossil fuel consumption
Permit users to turn off systems in part or whole	Reduce energy usage and fossil fuel consumption
Avoid high-energy materials	Reduce energy and preserves resources
Avoid high-energy processes	Reduce energy
Specify best-in-class energy efficiency components	Reduce energy usage and fossil fuel consumption
Insulate and/or use waste heat	Reduce losses/increases efficiency

9.6.4 Design for the Environment (DFE)

Design for the environment, like design for manufacturing or design for assembly, is a set of guidelines to help designers meet particular design goals. Often these guidelines are reduced to simple rules that aid understanding. However, behind these rules are observations and models that capture how the product can be expected to perform as the result of certain design decisions.

To a certain extent this whole chapter has been aimed at understanding how products and product decisions result in environmental loads. There can be, however, different environmental goals. For example, designing an automobile for lower fuel consumption may lead to using structural composite materials for weight reduction, whereas designing for recyclability would probably lead to the use of metals for the structural components. In this section we outline some of the generally agreed upon guidelines for two important environmental goals: reduced hydrocarbon fuel consumption and increased recyclability. These are summarized in Tables 9.24 and 9.25.

9.6.5 System-Level Observations

In this section we have presented an overview of engineering actions to lessen the impact of materials use, manufacturing, and design decisions on the environment. One of the goals of this section has been to identify the connections between a product life cycle and the associated environmental loads. To do this we have frequently normalized the environmentally sensitive parameters such as energy requirements or emissions by some measure of useful output such as the weight of the output, the economic activity, or, in some

Table 9.25 Recyclability guidelines [9.49, 78, 80]

Rating	Description or action	Reason or comment
Good	Product is reusable/remanufacturable	Extends life of product
Good	Materials in part are recyclable with a clearly defined technology and infrastructure	Most metals, some plastics in particular: PET & HDPE
Good	No toxic materials, or if present, clearly labeled and easy to remove	Avoid Pb, Hg, Cd
Good	Allow easy removal of materials, avoid adhesives and joining methods which cannot be reversed	Facilitates separation and sorting
Less good	Material is technically feasible to recycle but infrastructure to support recycling is not available	Most thermoplastics, some glass
Less good	Material is organic – can be used for energy recovery but cannot be recycled	thermoplastics, rubber, wood products
Avoid	Avoid mixtures which cause contamination, and painting and coatings which are difficult to remove	e.g., polyvinylchloride (PVC) in PET, Cu in steel, painted plastics
Avoid	Material has no known or very limited technology for recycling	Heated glass, fiberglass, thermoset plastics, composite materials

cases, just by the product itself. This scheme helps assign responsibility and allows us to track progress by enabling comparisons.

At the same time, however, by measuring the environmental load too narrowly there is a danger of missing the overall trend. One way of making this point is by writing the environmental impact in terms of several normalized parameters. For example, consider

$$\text{impact} = \text{population}\,\frac{\text{wealth}}{\text{person}}\,\frac{\text{impact}}{\text{wealth}}\,. \qquad (9.2)$$

This is a mathematical identity, known as the IPAT equation, which associates impact I, with three important elements: P for population, A for affluence, and T for technology. Our focus has been on the last term – impact/wealth (or impact/product etc.). Many variations on the IPAT equation are possible, for example $A =$ products/person, $T =$ impact/product, etc. It is the collection of the terms on the right-hand side that give the impact. Hence, a technology improvement could be offset by increases in population and/or wealth/person. This is unfortunate, but appears not to be in the domain of the engineer. If this was all there was to the story, the IPAT equation would be a neat way to subdivide responsibility. The implication is that, if engineers can improve the technology term, then they have done their job. The actual picture unfortunately is much more complicated, as technology improvements do not only improve the environment but also play an important role in stimulating the economy. In fact, relatively recent economic growth theories, pioneered by Nobel laureate *Robert Solow*, now give primary importance to technology change [9.81, 82]. Hence the very act of improving the performance of a product could, and often does, stimulate increased production and consumption of the product. Some versions of this effect are called the *rebound effect* or *Jevon's paradox*, after the 19th century economist who noted that more-efficient production and use of a resource (coal in his case), stimulated more consumption of the resource, not less [9.83].

In a similar vein, one could observe that taken as a whole labor-saving technological progress in developed countries has not led to less employment (but it has led to increased income). The general rule is that people respond to incentives, and all the incentives in a market economy point toward increasing investment and output rather than decreasing employment or resource use [9.82].

If society wants to reduce resource use, or emissions, or toxic waste, etc., it will need to provide the incentives, most likely through policy instruments, to do this. There are many successful examples to illustrate this point. The USA has reduced emissions of lead and sulfur dioxide, it has reduced the energy consumed by refrigerators, and the world has stabilized the levels of ozone in the upper atmosphere through implementation of the 1976 Montreal Protocol. Hence, the engineering actions described in this chapter should be taken in conjunction with a wider incentive and policy system that will preserve the engineering efficiency gains.

9.7 Failure Mode and Effect Analysis for Capital Goods

Failure mode and effect analysis (FMEA) [9.84, 85] is a method to recognize and eliminate mistakes or causes of faults during the product design process and in particular in the earliest stages.

This method was developed in 1963 by the National Aeronautics and Space Agency (NASA) within the Apollo mission to design products without design failures. This is especially important when products cannot be repaired easily, e.g., satellites or spaceships. The method was adopted later by the aviation industry, the automotive industry, in medicine and nuclear technology as well as by the armaments industry [9.85, 86]. Today this method is increasingly used in all fields of the development process of consumer and capital goods [9.84, 87–89].

With regard to the rework of the initially developed and applied approach see [9.89, 90].

9.7.1 General Innovations for the Application of FMEA

The complexity of many products due to their mechatronic character but also because of the simultaneously applied engineering concepts complicates the distinction between the three traditional types of FMEA: system FMEA, design FMEA, and process FMEA (Fig. 9.59). In addition many FMEA sessions mix all three fields [9.90].

Instead of this distinction a continuous form of FMEA therefore has to be implemented, which

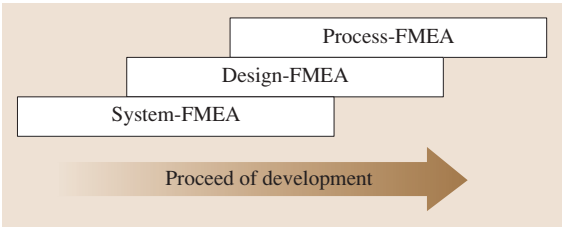


Fig. 9.59 FMEA separated in subprocesses (traditional form)

becomes more detailed with increasing knowledge through the product development process. As the design of a product proceeds and product knowledge grows, the danger of errors increases as does the need for their recognition and elimination. This leads to the concept of a continuous FMEA process (Fig. 9.60).

Such continuity produces the best results if the group of people involved stays the same. In practice, it turns out that it is already reasonable to apply competence in manufacturing or quality control in FMEA at the concept stage. This does not contradict the demand that specialists have to be consulted for specific questions. An additional benefit is gained from all team members becoming acquainted with the product.

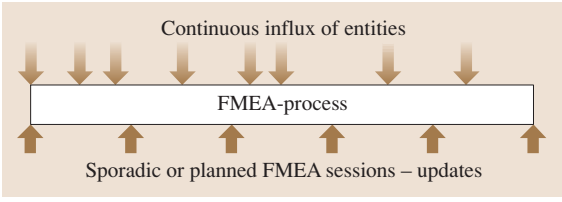


Fig. 9.60 FMEA as a continuous process

In general the early stages of the product development process and the stages during and after market launch are not covered by traditional FMEA. The early phases of the product development process (idea generation and market assessment) include many sources of errors that influence the success or failure (not only technically) of a new product.

The traditional form of FMEA does not take the costs of possible damage as well as the costs of avoiding this damage into account. To decide on appropriate trouble-shooting actions it is important to gain an economic understanding of FMEA and its consequences.

The new form, based on the original FMEA form but now including the economic point of view, is schematized and formalized in Fig. 9.61.

9.7.2 General Rules to Carry Out FMEA

The observation of FMEA sessions in companies shows that a clear separation into the three fields of system, design, and process is not performed, nor is it reasonable. This is because each failure affects the system, the design, and the process FMEA.

Since the system FMEA is carried out before the design FMEA this new influence into the system and his effects is not considered anymore. Only a completed process FMEA will discover the actual source of error, for example an unsuitable die-casting machine. Again the influence of the elimination of this failure mechanism is no longer considered in the system and design FMEAs.

Company project	**Failure mode and effect analysis – FMEA** Project- and lifecycle – FMEA									Name of functions / working principles / assemblies Ident-No.						
	Generating department:									Generated by: Date:						
Failurelocation Ident-No. Entity	Type of failure	Impact of failure	Cause of failure	Present status Means of control	A	B	E	K	RPZ	Remedial action	Improved status Realized action	A	B	E	K*	RPZ*
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17

Fig. 9.61 FMEA form with additional columns K and K^* to consider the cost influences for decision-making

For these reasons it is not reasonable to separate the three individual FMEA methods. It is more reasonable to carry them out simultaneously in order to benefit from synergy within the group.

The Prerequisites at the Beginning of a FMEA

The following prerequisites must be met absolutely at the beginning of an FMEA process:

1. FMEA participants that are competent and confirmed in their continuity [9.91, 92]. The definition of the *risk priority number* (RPZ) numbers (see Sect. 9.7.3) is done subjectively and thus randomly. Therefore the selection of the FMEA participants has to be done carefully in order to take into account different views and issues and to obtain a realistic RPZ. This problem is similar to that of the evaluation process [9.91, 92].
2. The product idea is defined and understood by all FMEA participants.
3. The market performance process of a new product is available.
4. A clear list of demands (requirements), tested for interrelations [9.93].
5. A use analysis [9.94].
6. A functional analysis, if possible as a functional diagram.
7. Heuristic context within the product development process. The methods involved between the individual components and units as well as external (referring to the usage of a product) and internal (working principles, effects and regularities) functions (Fig. 9.62) cannot be correctly represented in conventional forms, which will result in incorrect decisions.
8. Complete documentation of the state of development.

Improvement of Error Recognition

(Formal) FMEA with conventional forms [9.84] is in principle incomplete since a suitable inspection of completeness is not given. This takes effect on the methodic context and on the other hand on the accidental error recognition. Some ways of improving error recognition rates are

1. The experience of the FMEA participants
2. Integration of other specialists
3. Knowledge of or becoming acquainted with the product to be analyzed
4. Integration of customer-oriented departments and the customers themselves

	Examples		
Action	Operate	View	Hear
Function	Amplify	Shine	Ring
Principle	Hydraulic	Electric	Acoustic
Effect, physical principle	Static pressure	Ohm's law	Sound wave
Formula	$P_2 = P_1(A_2/A_1)$	$U = R\,I$	$E = J/(f\lambda)$

Fig. 9.62 Heuristic context within the product development process; the five stages of solution finding in the concept stage

5. Collection and analysis of mistakes and experiences recorded in databases
6. Study of the history of a product (test reports, complaint reports, etc.)
7. Concurrent application of heuristic support tools (e.g., *Goldfire*, Invention Machine, Boston)

Inclusion of all Secondary Fields of Development

The FMEA refers only to hardware, not to documentation, services, software logistics, test programs, manufacturing method, production plants, and machines. Measurement and checking facilities, experimental setups, and devices are not covered by any FMEA, although many errors could occur here. This situation only can be improved by including these new fields.

9.7.3 Procedure

The localization of risks concerning the lifetime of a product, starting from the first design concept to product use, is in principle always the same procedure, regardless of the product size. The principle and pro-

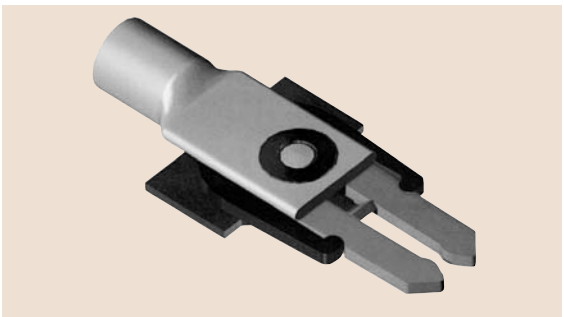


Fig. 9.63 Small product: plug-in contact (source: ABB, Baden)

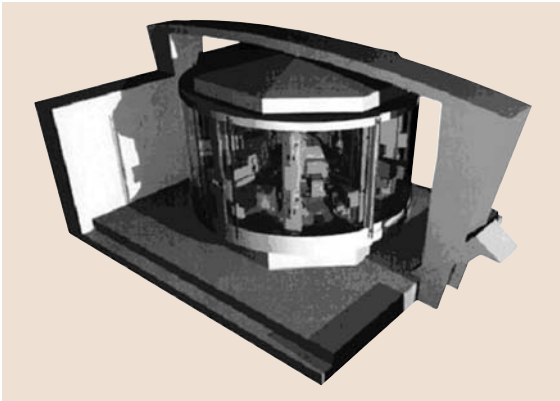


Fig. 9.64 Large product: machine tool (source: A. Breiing, MIKRON, Biel)

cedure is identical for a small product like a plug-in contact (Fig. 9.63) or a large technical system like a machine tool (Fig. 9.64).

Description of the project

Nomination of the FMEA moderator. The nomination is done with the agreement of the responsible project manager.

Formation of an FMEA team. The formation of the team (Table 9.26) is done by the project manager with the agreement of the FMEA moderator. Important is the visualization of the team members (Table 9.26). The departments should contribute at least one participant as a continuous FMEA team member.

Structuring of functions. The basis of FMEA is functional analysis of the investigated product. This implicates the specification of functional modules or submodules, for example as a hierarchical functional diagram (Fig. 9.65) or as a process-orientated diagram (Fig. 9.66).

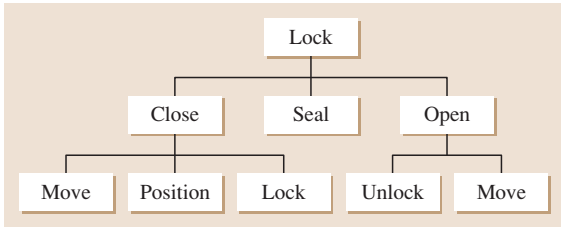


Fig. 9.65 Example of a hierarchical functional structure of an autoclave door [9.95]

Filling out the FMEA Form (Fig. 9.61).

Column 1. This column consists of all aspects of possible sources of error, as shown in Table 9.27. Since in the early stages of product development, in particular during idea generation and during the determination of the market performance profile, sources of error can already be detected, the corresponding entities have to be entered in this column (first block). Sources of error can also be detected in the development phase (second block), during market launch (third block), and in particular during the usage phase up to product disposal (fourth block).

Columns 2–4. These columns require an analysis of all possible potential failures, for example:

Possible errors in the planning stage.

- The idea of a new product cannot be realized due to physical, chemical, biological, ethical, legal or cultural constraints or restrictions
- The product idea is already realized by competitors
- The product idea is already protected by the competitors with patents, petty patents, and/or design patents
- There is no market potential and/or no market gap available
- The time is not ripe
- No foe image available or not relevant enough (for military development)

Table 9.26 Visualization of the team members

No.	Faculty/customer	Name, company
1	Project management	
2	Design, construction, and calculation	
3	Manufacturing preparation	
4	Manufacturing	
5	Quality control, simultaneously presentation of FMEA	
6	Purchase department	
7	Sales department	
8	Customer support (inauguration, maintenance, repair)	

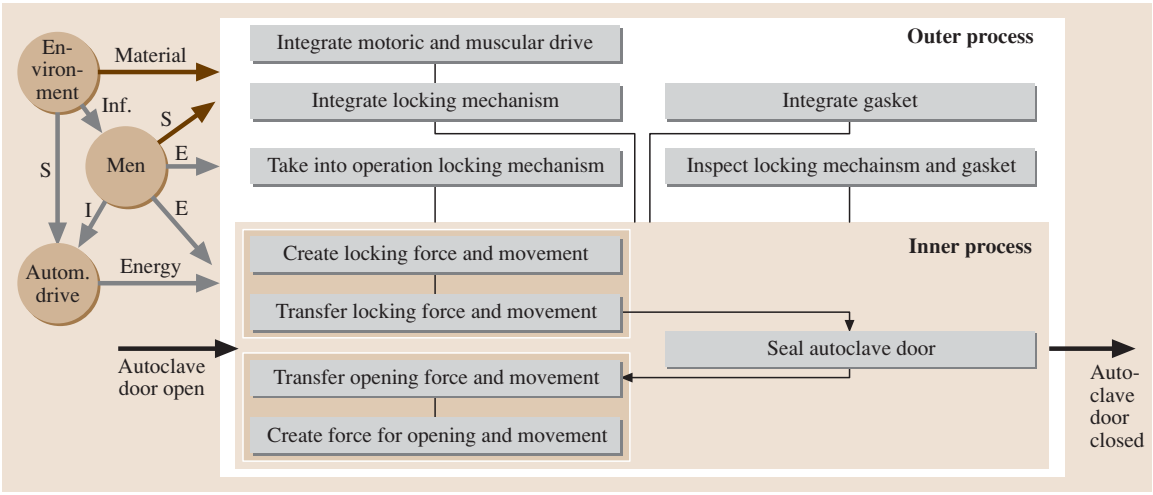


Fig. 9.66 Example of a process functional structure of an autoclave door [9.95], generated from Fig. 9.64

Table 9.27 FMEA form with entities over the complete life cycle of a product (an example)

Company **Project**	**Failure mode and effect analysis (FMEA)**			
Failure location Identification number Entity	Impact of failure	Type of failure	Failure cause	Means of control
1	2	3	4	5
Idea Market analysis Cost planning Time scheduling Strategy				
Inner functions Outer functions Module Assembly Component Manufacturing process				
Storage Transport Mounting Initiation				
Implementing Operation Controlling Maintenance Repair Recycling or liquidation				

- Rare resources
- No sustainable environment

Possible errors in the development process.
- Internal functions from a function analysis (hierarchical and/or process functional structure)
- External functions from a use analysis
- Assembly of the product structure
- Assembly/disassembly:
  - Assembly/disassembly operation
  - Assembly/disassembly regulations
  - Utilities (e.g., oil for the removal of bearings)
  - Assembly/disassembly devices
  - Control and/or measuring devices (e.g., torque spanner)
  - Additional means (e.g. hoisting devices)
- Components from the product structure and/or parts list
- Manufacturing:
  - Manufacturing processes (e.g., lapping)
  - Manufacturing operations (for manual manufacturing, e.g., deburring)
  - Manufacturing regulations
  - Manufacturing means and devices (e.g., drilling patterns)
  - Auxiliary manufacturing requirements (e.g., coolant)
  - Control and measuring devices (e.g., calipers)

The *outer function* takes into account the human being in his incompleteness as users during the product life. This must be included into the FMEA. The user is a component of the system; (s)he influences all processes, starting from idea generation, through development and usage until the end of the product life. The user causes failures and errors at all stages. With the integration of the available use analysis (Table 9.28) that covers all stages of a product's life [9.94] possible sources of error during assembly, maintenance, and repair are also considered (Table 9.27, fourth block).

In order to register these potential sources of error systematically the FMEA is expanded by the integration of the use analysis. These entities can be taken from the tabular recording of the man–machine interfaces.

Moreover each part, assembly, and product needs the following documentation:

Possible errors within the technical documentation. Check the:
- Manufacturing documents (drawings and lists)
- Documents for assembly/disassembly (e.g., instruction sheets)
- Calculation documents, such as:
  - Load assumptions
  - Verification of strength
  - Verification of deformation
  - Verification of stability (e.g., buckling, bending, stability)
- Balances such as:
  - Balance of performance
  - Balance of weight
  - Position of the centers of gravity
  - Balance of the moments of inertia
  - Balance of temperature
  - Balance of coolant
- Documents such as:
  - Manufacturing documents
  - Instruction sheets
  - Instructions for maintenance and service
  - Spare-part catalogue

Possible errors during market launch.
- Storage:
  - Activities
  - Apparatus (e.g., bearing block)
  - Racks, halls, stacks (e.g., storage rooms with air conditioning)
- Transport:
  - Activities
  - Apparatus (e.g., lifting gear)
  - Means of transportation (e.g., overhead crane)
- Mounting:
  - Activities
  - Mounting regulations and instructions
  - Apparatus (e.g., lifting gear)
  - Measuring and test equipment (e.g., theodolite)
- Initial operation:
  - Activities and tests in accordance with:
    - Instruction manual (Chapter *Initial setup*)
    - Checklist
    - Users' handbook
    - Maintenance instruction

Possible errors in the use stage and in decommissioning.
- Initial operation:
  - Activities in accordance with:
    - Instruction manual (Chapter *Initial setup*)
    - Checklist

Table 9.28 Use analysis of a nutcracker

No.	Subfunction first order	Man–product relationship: needed activities	Man–machine interface	Affiliated requirement	Required functions resp. possible carriers of function
1	Detection (of the object)	Nutcracker: Seek Ask for Find	Eye – object Sense of touch – object	Noticable design Recognizable design Shiny color	Color Contrast Grade of reflection
2	Transporting/ placing (the object)	Nutcracker: Grasp, lift, carry Put down Slacken	Hand – handle Hand – object body Eye – hand – handle	Little weight, ergonomic handle, handy surfaces, *crack protection* Stable stand	Lightweight construction Handle – leverage Corrugated surface Platform
3	Equipment (with a nut)	Open nutcracker Insert nut	Hand – object Hand – nut	Easy to equip Safely to equip	Trough Stop
4	Locating (of the nut)	Hold nut Press against stop Clamp nut	Finger – nut Hand – nut Eye – finger – nut	Easy to use Sure hold of nut Limit clamping force	Trough Clamping claw Vise
5	Produce (opening) force	Move leverage Hit against anvil Turn knob	Hand – leverage Fist – anvil Hand/finger – knob	Sure force insertion Notice finger/hand span	Handle Leverage
6	Guide force/ amplify force	Inner function	–	Force amount [N] Distance amount [mm] Limit force/distance	Leverage Spline Screw
7	Nut opens by: · pressure · effect of spline	Inner function	–	Selection of effective functionalities that can be cheaply realized	Thrust piece Splines Blade Clamping screw
8	Remove (result)	Remove cracked nut Remove nut and shell	Finger – crecked nut Hand – cracked nut Eye – finger – cracked nut	Easy to remove Sure to remove	Collecting pan, sack Basin Flap Opening
9	Cleaning	Hold nutcracker Shake out crack room Clean crack room	Hand – nutcracker Finger/hand – cleaning device	Easy to handle No unreachable corners Easy to clean surface	*Room design* Shape of surface Surface roughness

  - Maintenance history
  - Document
- Operating:
  - Activities in accordance with:
    - Instruction manual
    - Users' handbook
- Maintenance:
  - Activities and test in accordance with:
    - Instruction manual
    - Users' handbook
    - Service documentation
    - Maintenance history (e.g., exhaust gas document for a motor vehicle)
- Service:
  - Activities and test in accordance with:
    - Instruction manual
    - Users' handbook
    - Service manual and spare-part catalogue
    - Logistics documents (e.g., global workshop catalogue)

Column 5. Here the status of currently used measures for the prevention of failures and test procedures is entered. These entries are used to reduce the causes of failure in column 4 and to detect possible sources of error.

Columns 6–10. These columns are used to calculate the risk priority number (RPZ).

To decide on actions for trouble-shooting it is very important to obtain an economic perspective on FMEA and its consequences. In the case of extension of the FMEA by integrating a cost calculation, the form includes an additional column K for the costs. The factor K stands for the probable costs in the case of an error occurring.

The value of column 10 has to be calculated by

$$\mathrm{RPZ} = A \cdot B \cdot E \cdot K \,, \tag{9.3}$$

where: A = probability of a failure incidence

Unlikely	1
Very low	2
Low	3
Medium	4
High	5
Very high	6

B = consequences for the customer

Consequences not noticeable	1
Inconsiderable inconvenience for the customer	2
Small inconvenience for the customer	3
Inconvenience for the customer	4
Irritation of the customer	5
Possible loss of the customer	6

E = probability of failure detection

Very high	1
High	2
Medium	3
Low	4
Very low	5
Unlikely	6

K = probability costs in the case of a fulfilled error

No or negligible increase of costs	1
Little additional costs	2
Medium additional costs	3
High additional costs	4
Extremely high additional costs	5
Nonbudgeted costs	6

Instead of the absolute measure of 1–6 a range of values from 1–10 can also be chosen. This is only reasonable if detailed knowledge is available, for example within an FMEA at the end of the product development process.

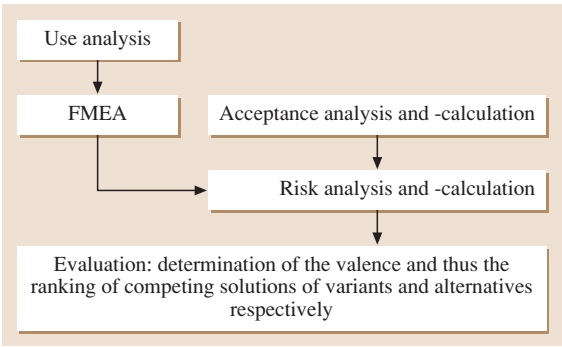


Fig. 9.67 Relationship between acceptance, risk and evaluation

A verbal description of the RPZ is:

Unacceptable high	1296–10 000
Medium	256–1295
Low	1–255

In practice RPZ failure values lower than 81 resp. 625 are not used to remove a failure as it is not profitable to do so.

Column 11. In this column the actions for trouble-shooting are entered.

Columns 12–17. Supposing that the recommended actions for trouble-shooting from column 11 are realized the actions are described in columns 12 and the RPZ* value is calculated again in columns 13–17. The factor K^* stands for the probable costs of the measure required to avoid or minimize a probable error. Both K and K^* have the same range of risk factors. The measure is accepted if the RPZ* is below the level described in the above.

The value of column 17 is calculated by

$$\mathrm{RPZ}^* = A \cdot B \cdot E \cdot K^* \,. \tag{9.4}$$

In practice failure values RPZ* that are lower than 81 resp. 625 are not used to remove a failure as it is not profitable to do so. The verbal description of RPZ* is like that of RPZ.

If RPZ* is higher than RPZ then the measure to avoid or to minimize a probable error has to be considered critically.

9.7.4 Further Use of FMEA Results

Within the concept and sketching phases several competing solutions or alternatives are available. In order to identify the best solution a ranking procedure is carried out. If the criterion *risk* is already part of this ranking procedure it is reasonable to use the sum of the RPZ or RPZ* values of each solution.

The relationship between acceptance, risk, and weighting is shown in Fig. 9.67.

References

9.1 K. Ehrlenspiel: *Integrierte Produktentwicklung: Denkabläufe, Methodeneinsatz, Zusammenarbeit*, 2nd edn. (Hanser, München 2002), in German

9.2 K. Roth: *Konstruieren mit Konstruktionskatalogen: Konstruktionslehre*, 3rd edn. (Springer, Berlin, Heidelberg 2001), in German

9.3 G. Pahl, W. Beitz, J. Feldhusen, K.-H. Grote: *Konstruktionslehre*, 7th edn. (Springer, Berlin 2007), in German

9.4 F. Kramer: *Innovative Produktpolitik* (Springer, Berlin 1988), in German

9.5 W. Rodenacker: *Methodisches Konstruieren*, 4th edn. (Springer, Berlin 1991), in German

9.6 R. Koller: *Konstruktionslehre für den Maschinenbau*, 4th edn. (Springer, Berlin 1998), in German

9.7 Dubbel: *Taschenbuch für den Maschinenbau*, 21st edn. (Springer, Berlin 2004), ed. by K.-H. Grote, J. Feldhusen, in German

9.8 H. Petra: Systematik, Erweiterung und Einschränkung von Lastausgleichslösungen für Standgetriebe mit zwei Leistungswegen. Ph.D. Thesis (TU München, München 1981), in German

9.9 DIN: *DIN 69910: Wertanalyse* (Beuth, Berlin 1987), in German

9.10 DIN: *Sachmerkmale, DIN 4000 – Anwendung in der Praxis* (Beuth, Berlin 2006), in German

9.11 DIN: *DIN 4000 (z.Zt. mit Entwürfen 163 Teile): Sachmerkmal-Leisten [für Norm- und Konstruktionsteile]* (Beuth, Berlin 2006), in German

9.12 DIN: *CAD-Normteiledatei nach DIN*, 3rd edn. (Beuth, Berlin 1984), in German

9.13 D. Krauser: *Methodik zur Merkmalbeschreibung technischer Gegenstände* (Beuth, Berlin 1986), in German

9.14 H. Czichos, M. Hennecke: *Hütte – Das Ingenieurwissen*, 33rd edn. (Springer, Berlin 2008), in German

9.15 H. Holliger-Uebersax: *Handbuch der allgemeinen Morphologie*, 4th edn. (MIZ, Zürich 1980), in German

9.16 J. Müller: *Grundlagen der systematischen Heuristik* (Dietz, Berlin 1970), in German

9.17 H.G. Schmidt: *Heuristische Methoden als Hilfen zur Entscheidungsfindung beim Konzipieren technischer Produkte* (TU Berlin, Berlin 1980), in German

9.18 V. Krick: *An Introduction to Engineering and Engineering Design*, 2nd edn. (Wiley, New York 1969)

9.19 R.K. Penny: Principles of engineering design, Postgrad. J. **46**, 344–349 (1970)

9.20 W.F. Daenzer: *Systems Engineering*, 6th edn. (Industrielle Organisation, Zürich 1989)

9.21 VDI: *VDI 2221: Methodik zum Entwickeln und Konstruieren technischer Systeme und Produkte* (VDI, Düsseldorf 1993), in German

9.22 F. Zwicky: *Entdecken, Erfinden, Forschen im morphologischen Weltbild*, 2nd edn. (Baeschlin, Glarus 1989), in German

9.23 H. Seeger: *Industrie-Designs* (Expert, Grafenau 1983), in German

9.24 H. Hertel: *Biologie und Technik* (Krausskopf, Mainz 1963), in German

9.25 P. Kerz: Konstruktionselemente und -prinzipien in Natur und Technik, Konstr. **39**, 474–478 (1987), in German

9.26 P. Kerz: Natürliche und technische Konstruktionen in Sandwichbauweise, Konstr. **40**, 41–47 (1988), in German

9.27 P. Kerz: Zugbeanspruchte Konstruktionen in Natur und Technik, Konstr. **40**, 277–284 (1988), in German

9.28 W. Kroy: Abbau von Kreativitätshemmungen in Organisationen. In: *Personal-Management in der industriellen Forschung und Entwicklung* (Heymanns, Köln 1984), in German

9.29 R. Kühnpast: Das System der selbsthelfenden Lösungen in der maschinenbaulichen Konstruktion. Ph.D. Thesis (TH Darmstadt, Darmstadt 1968), in German

9.30 K.-H. Habig: *Verschleiß und Härte von Werkstoffen* (Hanser, München 1980), in German

9.31 VDI: *VDI 2242 Blatt 1 und 2: Konstruieren ergonomiegerechter Erzeugnisse* (VDI, Düsseldorf 1986), in German

9.32 I. Klöcker: *Produktgestaltung* (Springer, Berlin 1981), in German

9.33 VDI: *VDI 2243 Blatt 1: Recyclingorientierte Produktentwicklung* (VDI, Düsseldorf 2002), in German

9.34 H. Meyer: *Recyclingorientierte Produktgestaltung* (VDI, Düsseldorf 1983), in German

9.35 M. Pourshirazi: *Recycling und Werkstoffsubstitution bei technischen Produkten als Beitrag zur Ressourcenschonung* (TU Berlin, Berlin 1987), in German

9.36 R.-D. Weege: *Recyclinggerechtes Konstruieren* (VDI, Düsseldorf 1981), in German

9.37 VDI: *VDI 2225 Blatt 1 und 2: Technisch-wirtschaftliches Konstruieren* (VDI, Düsseldorf 1998), in German

9.38 C. Zangemeister: *Nutzwertanalyse in der Systemtechnik*, 4th edn. (Wittemann, München 1976), in German

9.39 VDI: *VDI 2222: Methodisches Entwickeln von Lösungsprinzipien* (VDI, Düsseldorf 1997), in German

9.40 F.J. Gierse: *Funktionen und Funktionsstrukturen, zentrale Werkzeuge der Wertanalyse*, VDI-Berichte, Vol. 849 (VDI, Düsseldorf 1990), in German

9.41 V. Hubka: *Theorie Technischer Systeme: Grundlagen einer wissenschaftlichen Konstruktionslehre*, 2nd edn. (Springer, Berlin, Heidelberg 1984), in German

9.42 F. Hansen: *Konstruktionssystematik: Grundlagen für eine allgemeine Konstruktionslehre*, 2nd edn. (VEB Verlag Technik, Berlin 1965), in German

9.43 J. Rugenstein (Ed.): *Arbeitsblätter Konstruktionstechnik* (Technische Hochschule Magdeburg, Magdeburg 1978/1979), in German

9.44 F. Engelmann: *Produktplanung und Produktentwicklung in kleinen und mittleren Unternehmen* (Shaker, Aachen 1999), in German

9.45 Novespace: *Parabolic Flight Campaign with A300 ZERO-G User's Manual*, 5th edn. (Novespace, Paris 1999), www.novespace.com/VEnglish/Microgravity_a/man_vola/flightUserManual.htm

9.46 R. Bjärnemo: *Evaluation and Decision Techniques in the Enginnering Design Process* (Heurista, Zürich 1991)

9.47 A.F. Osborn: *Applied Imagination – Principles and Procedures of Creative Thinking* (Scribner, New York 1957)

9.48 J.O. Nriagu, J.M. Pacyna: Quantitative assessment of worldwide contamination of air, water and soils by trace metals, Nature **333**, 134–149 (1988)

9.49 T.E. Graedel, B.R. Allenby: *Design for Environment* (Prentice Hall, New York 1998)

9.50 W. Leontief: *Input-Output Economics*, 2nd edn. (Oxford Univ. Press, Oxford 1986)

9.51 C. Hendrickson, A. Horvath, S. Joshi, L. Lave: Economic input-output models for life-cycle assessment, Environ. Sci. Technol. **13**(4), 184A–191A (1998)

9.52 R. Miller, P. Blair: Input-output analysis: Foundations and extensions. In: *Environmental Input-Output Analysis* (Prentice Hall, New York 1985) pp. 236–260, Chap. 7

9.53 S. Joshi: Product environmental life-cycle assessment using input-output techniques, J. Ind. Ecol. **3**(2,3), 95–120 (2000)

9.54 S. Suh, G. Huppes: Methods for life cycle inventory of a product, J. Cleaner Prod. **13**, 687–697 (2005)

9.55 T.E. Graedel: *Streamlined Life-Cycle Assessment* (Prentice Hall, New York 1998)

9.56 Environmental Protection Agency: *EPA TRI 1998 Data Release Web Page* (EPA, Washington 1998), http://www.epa.gov/tri/

9.57 V. Smil: *Energies – An Illustrated Guide to the Biosphere and Civilization* (MIT Press, Cambridge 1999)

9.58 K.J. Martchek, E.S. Fisher, D. Klocko: Alcoa's worldwide life cycle information initiative, Proc. Total Life Cycle Conference – Land, Sea and Air Mobility, SAE Int. **P-339**, 121–125 (1998)

9.59 P.F. Chapman, F. Roberts: *Metals Resources and Energy* (Butterworth-Heinemann, London 1983)

9.60 E. Williams, R. Ayres, H. Heller: The 1.7 kg microchip: Energy and chemical use in the production of semiconductors, Environ. Sci. Technol. **36**(24), 5504–5510 (2002)

9.61 J. Dahmus, T. Gutowski: An environmental analysis of machining. In: *ASME Int. Mechanical Engineering Congress*, ed. by L. Yao (ASME, New York 2004)

9.62 S. Dalquist, T. Gutowski: Life cycle analysis of conventional manufacturing techniques: Sand casting. In: *ASME Int. Mechanical Engineering Congress*, ed. by L. Yao (ASME, New York 2004)

9.63 J. Sherman, B. Chin, P.D.T. Huibers, R. Garcia-Valls, T.A. Hatton: Solvent replacement for green processing, Environ. Health Persp. **106**, 253–271 (1998), Suppl. 1

9.64 V.M. Thomas, T.G. Spiro: The U.S. dioxin inventory: Are there missing sources?, Environ. Sci. Technol. **30**(2), 82A–85A (1996)

9.65 A. Grubler: *Technology and Global Change* (Cambridge Univ. Press, Cambrigde 1998)

9.66 J.L. Sullivan, R.L. Williams, S. Yester, E. Cobas-Flores, S.T. Chubbs, S.G. Hentges, S.D. Pomper: Life cycle inventory of a generic US family sedan – Overview of results USCAR AMP project. In: *SAE International 1998, Total Life Cycle Conference Proc.* (Society of Automotive Engineers, Warendale 1998) pp. 1–14, paper 982160

9.67 G.A. Keoleian, K. Kar, M.M. Manion, J.W. Bulkley: *Industrial Ecology of the Automobile: A Life Cycle Perspective* (Society of Automotive Engineers, Warrendale 1997)

9.68 H. Maclean, L. Lave: A life-cycle model of an automobile, Environ. Sci. Technol. **32**(13), 322A–329A (1998)

9.69 T.E. Graedel, B.R. Allenby: *Industrial Ecology and the Automobile* (Prentice Hall, New York 1998)

9.70 J.M. DeCicco, M. Thomas: A method for green rating of automobiles, J. Ind. Ecol. **3**(1), 55–75 (1999)

9.71 M.A. Weiss, J.B. Heywood, E.M. Drake, A. Schafer, F.F. AuYeung: *On the Road in 2020*, Energy Laboratory Report MIT EL 00-003 (MIT, Cambridge 2000)

9.72 R. Kuehr, E. Williams (Eds.): *Computers and the Environment Understanding and Managing their Impacts* (Kluwer Academic, Dordrecht 2004)

9.73 Microelectronics, Computer Technology Corporation: *Life Cycle Assessment of a Computer Workstation*, Report HVE-059-094 (MCC, Austin 1994)

9.74 K. Kawamoto, J. Koomey, B. Nordman, A. Meier: Electricity used by office equipment and network equipment in the U.S. In: *Conf. Energy Efficiency in Buildings* (EPA, Lawrence

Berkeley National Laboratory, Berkeley 2000), http://enduse.lbl.gov/projects/Indo/Puba.html
9.75 D. Cole: Energy consumption and personal computers. In: *Computers and the Environment: Understanding and Managing Their Impacts*, ed. by R. Kuehr, E. Williams (Kluwer Academic, Dordrecht 2003) pp. 131–159
9.76 E. Williams: Environmental impacts in the production of personal computers. In: *Computers and the Environment: Understanding and Managing Their Impacts*, ed. by R. Kuehr, E. Williams (Kluwer Academic, Dordrecht 2004) pp. 41–72
9.77 Environmental Protection Agency: *EPA egrid 2004* (EPA, Washington 1998), www.epa.gov/cleanenergy/egrid/index.htm
9.78 B. Bras: *Environmentally Conscious Design and Manufacture*, Lecture Notes ME, Vol. 4171 (Georgia Tech, Atlanta 2004), www.srl.gatech.edu/
9.79 K. Otto, K. Wood: *Product Design: Techniques in Reverse Engineering and New Product Development* (Pearson Education, Upper Saddle River 2001)
9.80 B. Metzger: *Design for Recycling: Influencing the Design Process at a Major Information Technology Company*, MS Thesis (MIT, Cambridge 2003)
9.81 R.M. Solow: Technical change and the aggregate production function, Rev. Econ. Statist. **39**, 312–320 (1957)
9.82 W. Easterly: *The Elusive Quest for Growth: Economists' Adventures and Misadventures in the Tropics* (MIT Press, Cambridge 2002)
9.83 W.S. Jevons: *The Coal Question: An Inquiry Concerning the Progress of the Nation, and the Probable Exhaustion of our Coal-mines*, Reprints of Economic Classics (Kelley, Fairfield 1906)
9.84 DIN: *DIN 25448: Ausfalleffekt-Analyse* (Beuth, Berlin 1978), in German
9.85 *Procedures for Performing a Failure Mode, Effects and Criticality Analysis (FMECA)*, MIL Std. 1629A (Military Standard, Washington 1980)
9.86 P. Conrad, P.E. Hedderich: *Navy Proactive Maintenance* (US Navy, Washington 2000)
9.87 N. Berens: *Anwendung der FMEA in Entwicklung und Produktion* (Verlag Moderne Industrie, Landsberg 1989), in German
9.88 C.H. Kepner, B.B. Tregoe: *Entscheidungen vorbereiten und richtig treffen* (Verlag Moderne Industrie, Landsberg 1988), in German
9.89 M. Schubert: *FMEA – Fehlermöglichkeits- und Einflußanalyse* (Deutsche Gesellschaft für Qualität, Frankfurt 1993), in German
9.90 A. Breiing: *Die FMEA in sinnvoller Form für Investitionsgüter* (Institut für Mechanische Systeme, ETH Zürich 2003), in German
9.91 A. Breiing: The evaluators influence on the results of evaluation, MCE 2000, Neukirchen (2000)
9.92 A. Breiing: Who evaluate the evaluators?, Int. Conf. Computer Integrated Manufacturing (Zakopane 2001)
9.93 A. Breiing, R. Knosala: *Bewerten Technischer Systeme* (Springer, Berlin, Heidelberg 1997), in German
9.94 A. Breiing: *Vertiefungsvorlesung Produkte-Design* (Institut für Mechanische Systeme, ETH Zürich 2000), in German
9.95 A. Breiing, M. Flemming: *Theorie und Methoden des Konstruierens* (Springer, Berlin, Heidelberg 1993), in German

10. Piston Machines

Vince Piacenti, Helmut Tschoeke, Jon H. Van Gerpen

Piston machines are the most used power and work machines in the mechanical engineering industry. The piston machines are divided in so-called reciprocating and rotary piston machines. With the first one a reciprocating motion is transformed to a rotary motion in the case of the power machine and conversely in the case of the working machine. Today rotary piston machines are almost exclusively used as work machines. Important innovations and intensive researches are practiced particularly for the use of the piston machines as an internal combustion engine. Therefore the mixture formation and the combustion process, with their consequences in terms of emission and fuel-consumption are int the center of attention.

10.1 Foundations of Piston Machines

10.1.1 Definitions

Piston machines employ a moving displacer (also called a piston) to convert a medium's potential energy into kinetic energy or vice versa, i. e., they use the movement of the displacer to increase the energy content of the medium. This occurs in a working chamber that can be altered by the displacer motion.

In piston machines, the moving displacer effects both the charge cycle (filling and draining of the medium) and the work cycle (expansion and compression). The characteristic mode of operation for piston machines is a self-contained working chamber that varies periodically due the piston's movement.

Piston machines can be classified according to their method of operation, the piston motion and the medium used (Fig. 10.1). When piston machines are classified according to their method of operation then power and work machines are differentiated.

A power machine converts a medium's potential energy into mechanical energy. Power machines include engines (pneumatic engine, hydraulic motor) and thermal machines (steam engines, combustion engines). Work machines on the other hand utilize mechanical energy to increase the energy of the medium being

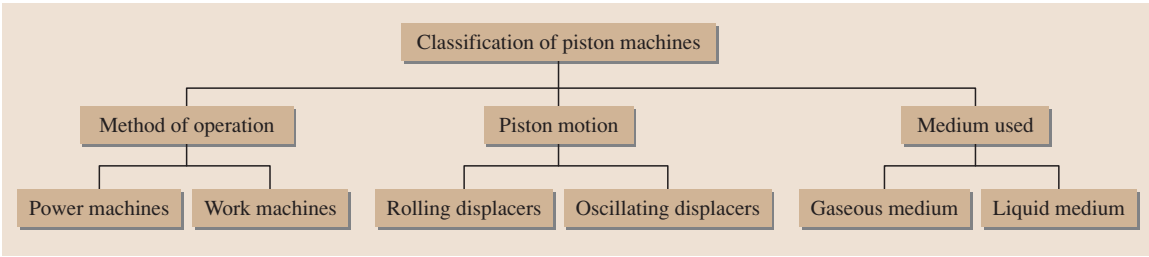


Fig. 10.1 Classification of piston machines (after [10.1])

conveyed. Work machines include compressors and pumps.

Rotating and oscillating displacer motions are other potential classification.

In addition, piston machines can be classified according to the medium used, which can be gaseous or liquid.

Reciprocating Machines

When a piston machine's oscillating displacer executes a linear motion, it is called a reciprocating machine (Fig. 10.2). A cylindrical displacer that moves between two end positions, top and bottom dead center (TDC and BDC), is characteristic. The working chamber is calculated from the piston diameter and the distance between the two dead centers, the so-called stroke s (see *Working Chamber* below), and is therefore also referred to as the displacement.

Rotary Piston Machines

Machines of this type are characterized by a rotating displacer. The medium can flow axially or radially to the piston axis. An axial direction of flow is found, for instance, in screw-type compressors, screw pumps or eccentric screw pumps. Rotary piston compressors, Roots blowers and gear pumps are examples in which the medium flows radially to the piston axis. In Wankel engines (Fig. 10.3), however, the medium flows axially as well as radially, depending on the design.

Cylinder Configuration

Differently configuring the interrelation of the cylinders makes it possible to produce various types of piston machines.

The inline and V configuration in particular and the boxer variant to a lesser degree are choices for combustion engines. The V design is compact and produces high output per unit volume. Its compact design makes manufacturing it more complex and expensive though. Moreover, accessibility is impaired, as a result of which

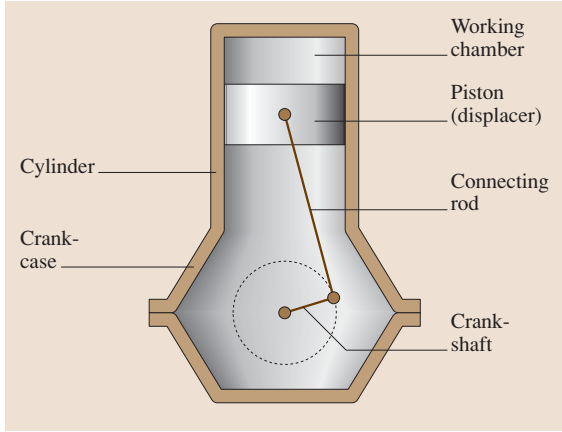


Fig. 10.2 Reciprocating machine

the time and effort required for maintenance increase along with maintenance costs. The advantages of the boxer engine are its overall length and height, though at the expense of width. The W configuration is frequently employed for piston compressors, yet relatively infrequently for combustion engines. The radial configuration is considered obsolete. It was used early on for

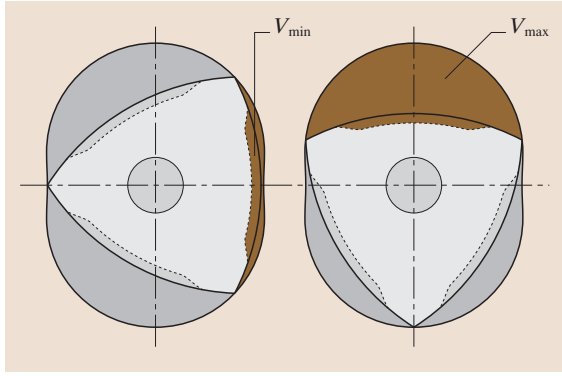


Fig. 10.3 Wankel engine as an example of a rotary piston machine (after [10.2])

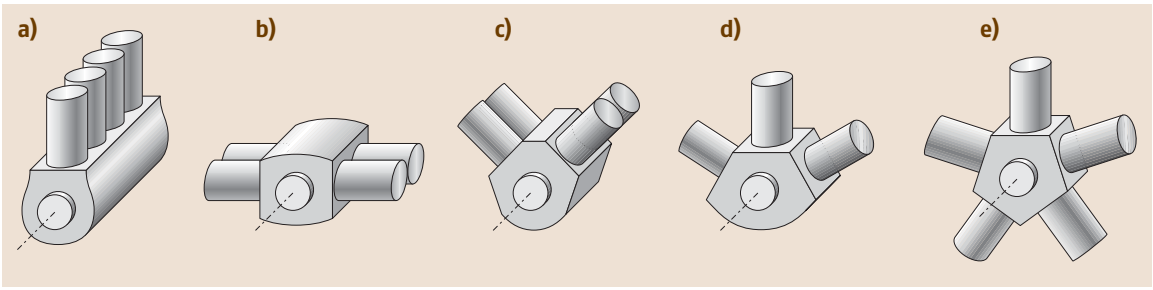


Fig. 10.4a–e Cylinder configurations (after [10.1]) **(a)** in-line machine **(b)** opposed-cylinder machine **(c)** V machine **(d)** W machine **(e)** radial machine

aircraft engines because of the good air cooling resulting from the design and the minimal axial extension.

A multi-cylinder design helps increase machine performance for combustion engines. In compressors, multistage compression is achieved by additionally using varying piston diameters. Increasing the number of cylinders boosts running smoothness. A larger number of cylinders adversely affects production and maintenance costs. Moreover, a more-complex design increases susceptibility to failure. Cylinder configuration, unit size, and the stiffness of the crankshaft can limit the number of cylinders.

Working Chamber

The displacer's movement causes the actual working chamber volume V_a to vary between the volume limits V_{min} and V_{max}. V_{min} can be a design-related clearance volume V_S or, in the case of combustion engines, a compression volume V_c contingent on the working process. Thus $V_{min} \leq V_a \leq V_{max}$ applies.

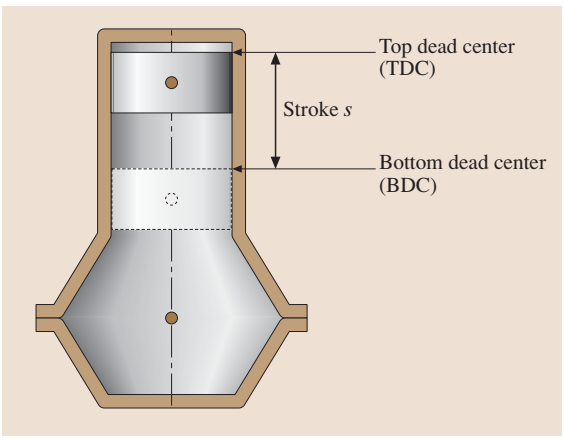


Fig. 10.5 The stroke s corresponds to the distance between top and bottom dead center

In a reciprocating machine, the maximum piston displacement V_A or swept volume V_h corresponds to the volume resulting from the product of the piston surface A_p and the stroke s (10.1). The stroke corresponds to the distance that the piston covers between the top and bottom dead centers (Fig. 10.5)

$$V_A = V_h = A_p s = s \cdot \pi \cdot D^2/4 \,. \qquad (10.1)$$

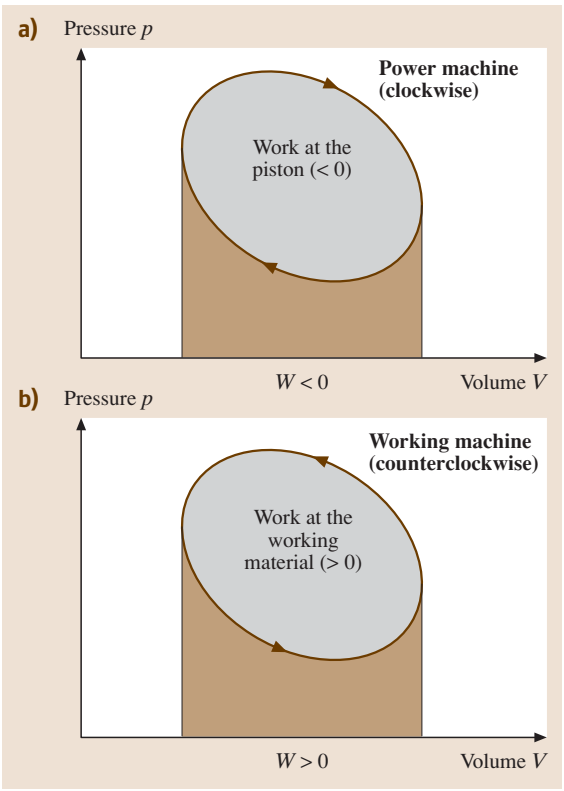


Fig. 10.6 p–V diagrams of power and work machines (after [10.3])

Defining a cylinder's volume (corresponding to the maximum volume) requires the incorporation of the minimum volume

$$V_{cyl} = V_{max} = V_h + V_{min} \,. \tag{10.2}$$

For machines with a multi-cylinder design, the total piston swept volume V_H follows from the number of cylinders z and the swept piston volume V_h

$$V_H = zV_h = zA_p s \,. \tag{10.3}$$

Using (10.2), the total working chamber of a combustion engine is calculated as

$$V_{working\text{-}chamber} = (V_h + V_{min}) \,. \tag{10.4}$$

10.1.2 Ideal and Real Piston Machines

The Ideal Combustion Cycle

A cycle is a succession of a material's changes of state until it returns to its initial state. The cycle serves as the foundation for evaluating the thermodynamics of processes.

The p–V diagram is used to represent the development of cycles (Fig. 10.6). Its enclosed area corresponds to the cycle work W.

Figure 10.6 is a general representation of the development of the cycles of a power and a work machine. The different direction of rotation of each process and the resultant sign of the cycle work are characteristic here. Since the work done by the machine's piston can be utilized (effective work), it is defined negatively, while the work expended on the working material is defined positively. The cycle work is calculated from the total of both values and corresponds to the area within the circle

$$W = W_{at\text{-}the\text{-}piston} + W_{at\text{-}the\text{-}working\text{-}material} \,, \tag{10.5}$$

$$W = \oint V\mathrm{d}p = -\oint p\mathrm{d}V \,. \tag{10.6}$$

A clockwise cycle profile with negative cycle work results for the power machine, while a counterclockwise cycle profile with positive cycle work results for the work machine.

Three examples of cycles based on the Carnot process (Chap. 4) are described below and are referred to as reference cycles. They are subject to the following premises:

1. State changes are infinitely slow.
2. The working chamber is adiabatically and hermetically sealed.

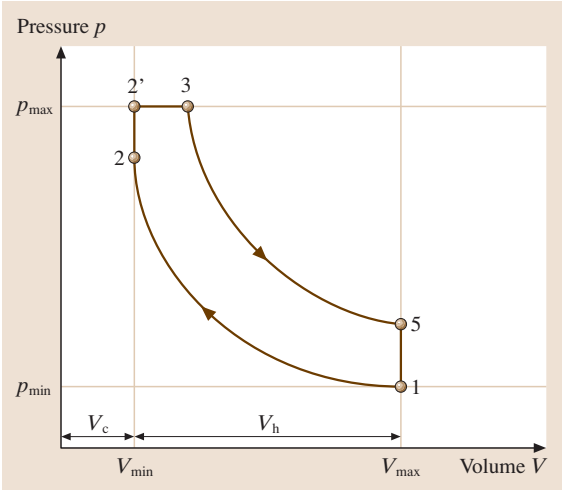


Fig. 10.7 p–V diagram of a combustion engine (Seiliger process) (after [10.4])

3. Fluid is exchanged without any change in the state variables (mass, pressure, temperature).

Combustion Engines. The Seiliger process constitutes a reference process for diesel engines. Adiabatic compression causes the pressure to rise until it is slightly below the maximum cylinder pressure. The internal combustion of a fuel mass is followed by an isochoric and isobaric input of heat, as a result of which the maximum cylinder pressure is reached. This is followed by an isentropic expansion up to the initial volume. The

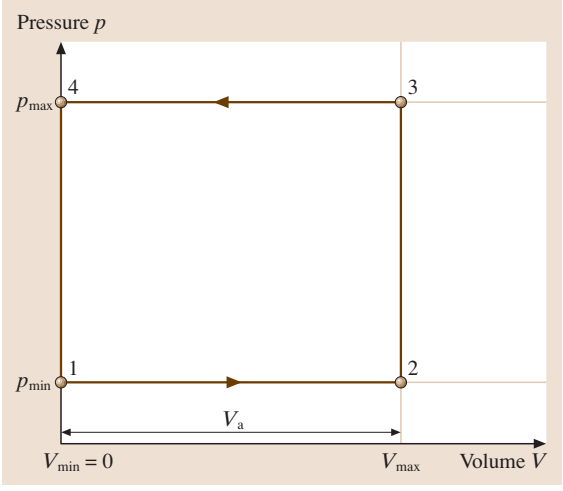


Fig. 10.8 p–V diagram of a positive displacement pump (after [10.4])

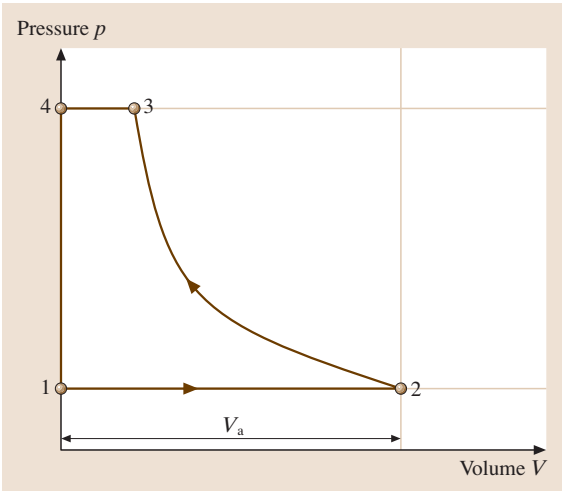


Fig. 10.9 p–V diagram of a positive displacement compressor (after [10.4])

initial state has been reached again when the heat extraction is isochoric.

Positive Displacement Pump. In an initial step, the working chamber is filled with an incompressible fluid. When the maximum volume V_{max} is reached, the pressure climbs isochorically to the maximum pressure p_{max}. The fluid is expelled at the maximum pressure. When the working chamber has been completely discharged, the pressure falls isochorically, returning to the initial state.

Positive Displacement Compressor. Analogous to the positive displacement pump, the working chamber is filled isobarically with a fluid up to the volume limit. When the positive displacement compressor's cycle is ideal, the low compression work causes the pressure to rise isothermally. When the maximum pressure has been reached, the fluid becomes isobaric and is expelled isochorically until it has been completely discharged.

Real Machines

While lossless state changes are assumed for an ideal combustion cycle, losses in the real machine occur as irreversible subcycles when the state changes. Infinitely slow (quasistatic) state changes would be needed to prevent irreversible cycles.

Deviations from the ideal cycle occurring in real machines are caused by:

1. Wall heat losses, i. e., heat exchange with the system boundaries (e.g., with the working chamber's walls and the piston)

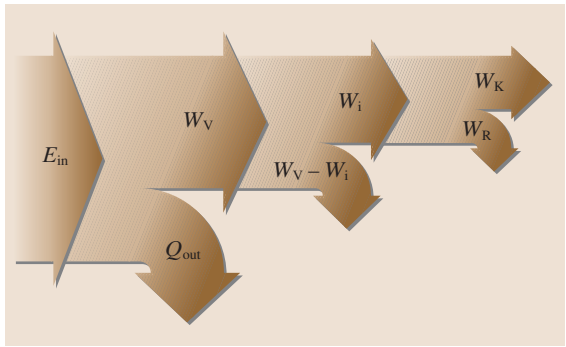


Fig. 10.10 Schematic diagram of a real power machine's losses (combustion engine)

2. Irreversibilities when states change; changing material values during the cycle
3. Losses when energy is transferred (e.g., friction in the machine, flow losses)
4. Leaks because the working chamber is not absolutely sealed

Efficiency

The aforementioned deviations are expressed by the efficiency, which represents the relationship between utility and effort. For heat engines, the total effort corresponds to the total quantity of input heat, while the utility is the outcome of an energy conversion process.

Power Machine. Figure 10.10 is a schematic diagram of a real power machine's losses.

Heat exchange with the system boundaries represents the greatest percentage loss of input energy. The relationship between the cycle energy of an ideal power machine's reference cycle and the input energy is expressed with the efficiency of the reference cycle and referred to as the efficiency of the ideal machine

$$\eta_v = W_v / E_{in} \, . \tag{10.7}$$

The irreversibilities of a cycle's state changes cause further losses. The related efficiency specifies the indicated cycle's approximation of the reference cycle and is referred to as a cycle's efficiency

$$\eta_g = W_i / W_v \, . \tag{10.8}$$

The work against friction W_R subsumes the losses occurring during the transfer of energy. The relationship between the effective work or the useful work that can be extracted at the clutch W_K and the indicated work is referred to as the mechanical efficiency

$$\eta_m = W_K / W_i \, . \tag{10.9}$$

Consequently, a real power machine's coupling efficiency is calculated from the efficiency chain

$$\eta_K = \eta_v \eta_g \eta_m = (W_v/E_{in})(W_i/W_v)(W_K/W_i) = W_K/E_{in} . \quad (10.10)$$

This efficiency represents all losses occurring in the power machine, which leads to the output useful work (clutch work) W_K.

Work Machine. Analogous to the power machine, Fig. 10.11 is a schematic diagram of a real work machine's losses.

The following relationships likewise apply to the mechanical efficiency

$$\eta_m = W_i/W_K , \quad (10.11)$$

internal efficiency

$$\eta_g = W_V/W_i \quad (10.12)$$

and system efficiency

$$\eta_A = E_{benefit}/W_V . \quad (10.13)$$

Consequently, a work machine's total efficiency is calculated from the chain of efficiencies

$$\eta = \eta_m \eta_g \eta_A = (W_i/W_K)(W_V/W_i)(E_{benefit}/W_V) = E_{benefit}/W_K . \quad (10.14)$$

Specific Energy

The relationship between the piston machine's output energy W_e and the corresponding piston swept volume is referred to as the volume-specific energy w_e

$$w_e = W_e/V_H . \quad (10.15)$$

In combustion engines, the unit of the volume-specific energy w_e is denoted as kJ/dm^3. Work machines' mass-specific energy w' is obtained from the relationship of the requisite work energy W_a and the mass of the fluid delivered per working cycle m_f. It is specified, for instance, in kJ/kg

$$w' = W_a/m_f . \quad (10.16)$$

The more losses that occur in the work machine, the more energy is needed to deliver the same amount of fluid. Consequently, the mass-specific work also increases as losses occur.

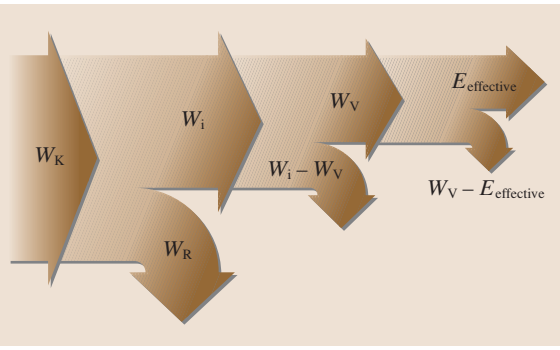


Fig. 10.11 Schematic diagram of a real work machine's losses

Power

A machine's power corresponds to the energy output (power machine) or input (work machine) per unit time.

$$P = \frac{W}{t} . \quad (10.17)$$

Moreover, the power can be calculated using the product of the torque M_d and angular velocity ω

$$P = M_d \omega = M_d 2\,\pi n , \quad (10.18)$$

where ω is the angular velocity and n is the engine speed or with the aid of the cycle energy W_K

$$P = \frac{W_K}{t_{ASP}} = W_K f_{ASP} , \quad (10.19)$$

where $t_{ASP} = i/2n$ is the time for one working cycle and $f_{ASP} = 1/t_{ASP}$ is the working cycle frequency; i is the number of cycles or piston strokes per working cycle. Expressing this using the mass flow $\dot{m}$ and the mass delivered per working cycle m_{ASP} leads to

$$P = \frac{W_K}{m_{ASP}} \dot{m} . \quad (10.20)$$

10.1.3 Reciprocating Machines

Types of Transmissions

Different types of transmissions transmit energy between piston and crankshaft, e.g., crankshaft drive, swash plate drive, cam drive and eccentric drive. The simple engineering and dimensioning of a reciprocating machine's important parameters (e.g., stroke and compression) are the reason why a crankshaft drive is frequently used as the transmission.

The connecting rod transmits energy between the piston and crankshaft (Fig. 10.2). In the process, the piston's oscillating motion is converted into the crankshaft's rotating motion in order to improve the further transmission of energy. Accordingly, the motion is transformed conversely when energy is transmitted from the crankshaft to the piston. The crankshaft drive can be constructed as a plunger piston or crosshead version. The crosshead serves to relieve the piston from

the lateral piston force, particularly in large piston machines.

Kinematics

The piston executes a motion characterized by accelerations and decelerations between the two dead centers. The piston's linear, irregular motion at the crankshaft is converted by the connecting rod into a rotating motion with constant angular velocity. The instantaneous distance covered by the piston, related to the top dead center, is designated the piston travel x (Fig. 10.12). The maximum piston travel, i. e., the distance between the two dead centers, is defined as the stroke s. This corresponds to twice the distance between crankshaft journal and shaft journal (crank radius r).

Piston travel can be specified as a function of the piston angle α

$$x = l + r - (l\cos\beta + r\cos\alpha)\,. \tag{10.21}$$

By substituting the connection rod angular travel β with α by using the relation

$$l\sin\beta = r\sin\alpha \tag{10.22}$$

and the connecting rod ratio

$$\lambda = r/l\,, \tag{10.23}$$

the formula for piston travel is obtained as

$$x = r(1-\cos\alpha) + l(1-\sqrt{1-\lambda^2\sin^2\alpha})\,, \tag{10.24}$$

where $\sin\beta = \lambda\sin\alpha$ and $\cos\beta = \sqrt{1-\lambda^2\sin^2\alpha}$.

Expanding the expression under the root into a Taylor polynomial produces a complex equation. In practice, only the first two orders of the Taylor polynomial are applied since they already yield sufficiently precise results for the piston travel x

$$x \approx r\left(1-\cos\alpha + \frac{\lambda}{2}\sin^2\alpha\right)\,. \tag{10.25}$$

Piston speed can be stated as an average or instantaneous speed.

The average piston speed is calculated based on the simple relation between distance covered and the necessary time expended. Twice the piston stroke is covered during one revolution. This yields the formula for the average piston speed

$$c_\mathrm{m} = \frac{2s}{\frac{1}{n}} = 2sn\,. \tag{10.26}$$

The following numerical value equation is also frequently applied

$$c_\mathrm{m} = \frac{sn}{30}\,. \tag{10.27}$$

The unit of revolution has to be specified in 1/min or min^{-1} and rpm respectively in order to obtain the average piston speed in m/s.

If the actual piston speed v is required, piston travel must be differentiated with respect to the time t, remembering that piston travel is only a function of the piston angle

$$v = \frac{\mathrm{d}s}{\mathrm{d}t} = \frac{\mathrm{d}s}{\mathrm{d}\alpha}\frac{\mathrm{d}\alpha}{\mathrm{d}t}\,, \tag{10.28}$$

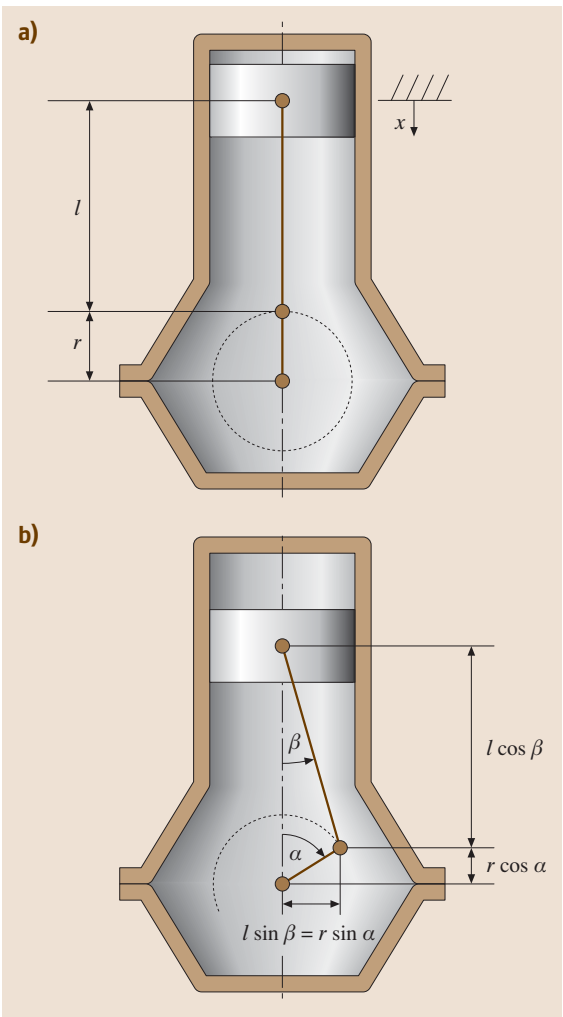


Fig. 10.12 Piston travel x and crankshaft drive geometry

where $d\alpha/dt$ is the crankshaft angular velocity ω.

Differentiating (10.25) yields an approximate equation for the piston speed:

$$v \approx \omega r \left[\sin\alpha + \frac{\lambda}{2} \sin(2\alpha) \right] . \tag{10.29}$$

Analogous to the piston speed, an approximate value can be derived for the piston acceleration

$$a \approx \omega^2 r [\cos\alpha + \lambda \cos(2\alpha)] . \tag{10.30}$$

Equations (10.25), (10.29) and (10.30) do not yield any precise results when speeds are high (e.g., racing engines). The exact piston travel and the corresponding derivatives for speed and acceleration would have to be used for this application.

Forces

Fluid Forces. The state change of the fluid enclosed in the working chamber induced by the oscillating motion of the displacer causes a force that is dependent on the piston surface A_K and the fluid pressure p_F exerted on the piston. Allowing for the atmospheric pressure p_0, the fluid force F_F of a single-action piston is calculated with the following equation

$$F_F = A_K (p_F - p_0) . \tag{10.31}$$

An equal force acts on the cylinder cover. The bolted connections, e.g., between the cylinder cover and the housing, create a closed linkage so that the forces in the machine balance (Fig. 10.13).

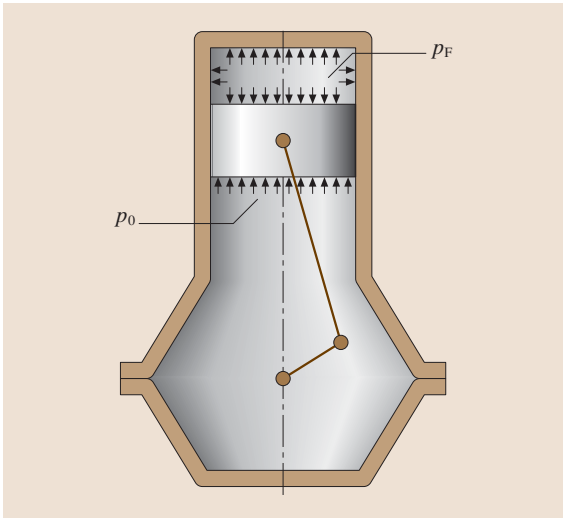


Fig. 10.13 Pressures on the piston

Inertial Forces. The irregular motions of the translationally moving structural parts cause periodic inertial forces to arise, which act as oscillators. The centripetal acceleration generates rotary inertial forces on the connecting rod and crankshaft suspensions. Therefore they have to be balanced. This can be easily and completely done for rotary piston machines, but only partially or with considerable effort for reciprocating machines. The inertial forces are classified as rotating and oscillating based on the different types of motion of the transmission components.

The rotating inertial forces F_{mr} are calculated with the formula

$$F_{mr} = m_r r \omega^2 . \tag{10.32}$$

The rotating masses m_r are concentrated in the crankshaft journal at the distance r from the center of rotation and consist of the masses of the crankshaft journal m_z, the crankshaft web (converted to the crank radius), and the rotating portion of the connecting rod (10.36).

The oscillating and rotating motion of the connecting rod cause the mass to be distributed to two points

$$m_{Pl} = m_{Plo} + m_{Plr} , \tag{10.33}$$

where m_{Pl} is the connecting rod mass, m_{Plo} is the oscillating mass fraction, m_{Plr} is the rotating mass fraction, a and b are the distances to the connecting rod center, and l is the distance between the two connecting rod eyes.

The mass m_{Plo} is located in the piston pin and is only involved in the oscillating motion, while m_{Plr} on the crankshaft journal executes a purely rotary motion. Since the distances from the center and the total moment of inertia are maintained, the masses are calculated based on

$$m_{Plo} = \frac{m_{Pl} a}{l} \tag{10.34a}$$

and

$$m_{Plr} = \frac{m_{Pl} b}{l} . \tag{10.34b}$$

As a result, a substitute connecting rod is obtained, the dynamic behavior of which is approximately comparable with a real connecting rod.

Since the center of the web masses m_W does not lie at the center of the crankshaft axis, the masses must be converted to the crank radius. Using the distance x between the center of the web masses and the crankshaft axis as well as the crank radius, the reduced mass m_{WE} is calculated as

$$m_{WE} = m_W \frac{x}{r} . \tag{10.35}$$

Taking the preceding observations and allowing for two web masses, the following result ensues for the rotating inertial forces of the reciprocating machine

$$F_{mr} = (m_Z + m_{Plr} + 2m_{WE})r\omega^2. \quad (10.36)$$

The irregular piston motion generates oscillating inertial forces F_{mo} as a function of the piston acceleration a

$$F_{mo} = m_o a\ , \quad (10.37)$$

where $m_o = m_K + m_{Plo}$ are the oscillating masses.

Taking the piston acceleration (10.30), the following ensues for the oscillating inertial force

$$F_{mo} \approx m_o\omega^2 r[\cos\alpha + \lambda\cos(2\alpha)]\ . \quad (10.38)$$

It is divided into oscillating inertial forces of first and second order (I and II)

$$F_{moI} = m_o\omega^2 r\cos\alpha = F_I\cos\alpha\ , \quad (10.39a)$$

$$F_{moII} = m_o\omega^2 r\lambda\cos(2\alpha) = F_{II}\cos(2\alpha)\ . \quad (10.39b)$$

The frequency of the first-order forces of inertia corresponds to the crankshaft speed, while the second-order forces of inertia change as the crankshaft speed doubles.

The influence of the stroke/connecting rod ratio λ (10.23) makes the amplitude of the second-order forces of inertia less than that of the first order. The extreme values for F_{moI} are in the top and bottom dead center, i. e., at $\alpha = 0°$ and $\alpha = 180°$. Since the frequency of F_{moII} is doubled, the maximum amplitudes are at $\alpha = 0, 90, 180$ and $270°$. This yields the following for the oscillating inertial forces at top dead center

$$F_{mo} = F_{moI} + F_{moII} \quad (10.40)$$

and the following at bottom dead center

$$F_{mo} = F_{moI} - F_{moII}\ . \quad (10.41)$$

Since there is no closed linkage, the oscillating inertial forces cause the crankcase to oscillate and thus exerts a load on the external engine suspension.

Forces on the Transmission. The piston force F_K consists of the fluid force F_F and the oscillating inertial force F_{os}. It is broken down into a lateral force F_N and a connecting rod force F_S as follows (Fig. 10.15)

$$F_N = F_K\tan\beta\ , \quad (10.42a)$$

$$F_S = \frac{F_K}{\cos\beta}\ . \quad (10.42b)$$

The cylinder wall absorbs the normal force. The connecting rod force divides at the crankshaft journal into a radial and a tangential component (F_R and F_T)

$$F_R = F_S\cos(\alpha+\beta) = F_K\frac{\cos(\alpha+\beta)}{\cos\beta}\ , \quad (10.43a)$$

$$F_T = F_S\sin(\alpha+\beta) = F_K\frac{\sin(\alpha+\beta)}{\cos\beta}\ . \quad (10.43b)$$

The tangential force and the crank radius produce the torque

$$M_d = F_T r\ . \quad (10.44)$$

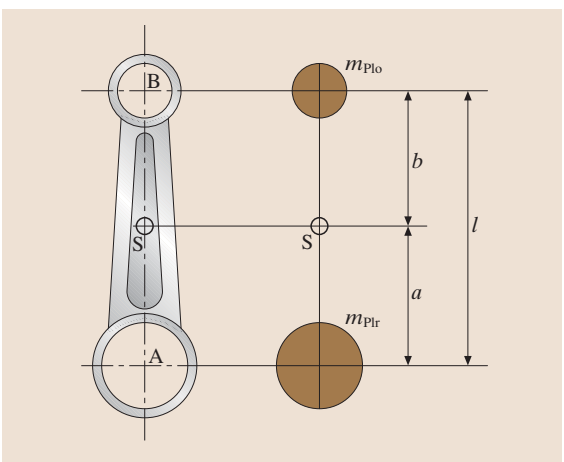


Fig. 10.14 Distribution of the connecting rod mass (after [10.5])

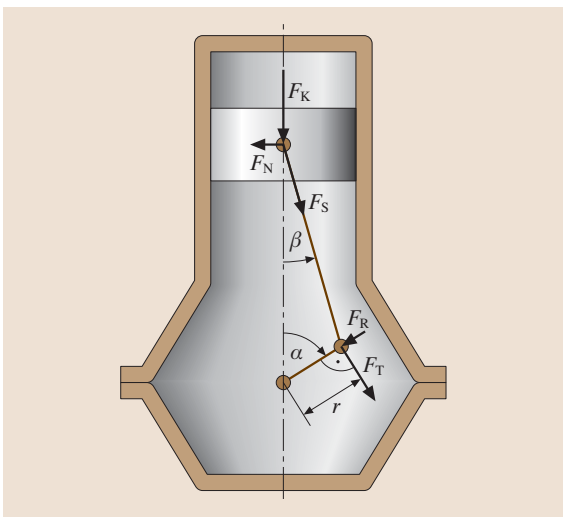


Fig. 10.15 Forces on the transmission

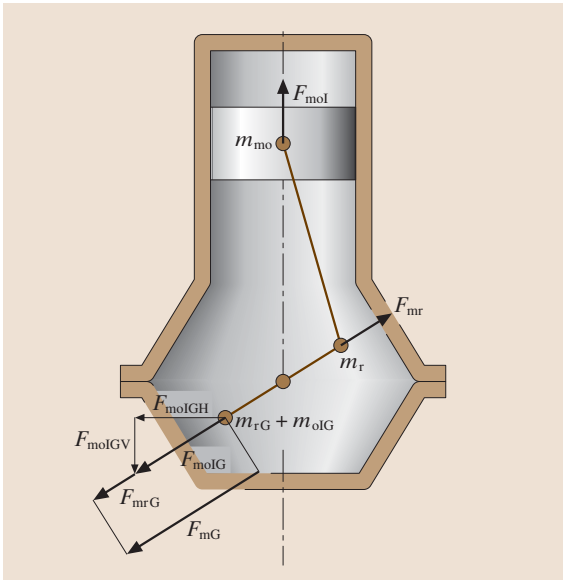


Fig. 10.16 Counterweight and forces

The direction of the force F_T corresponds to the direction of rotation. The normal force generates the corresponding reaction torque, which acts as a tilting moment on the housing and must be absorbed by the housing suspension.

Mass Balancing

The periodically varying inertial forces and moments of the transmission's moving points can cause oscillations, which have an effect on the engine suspension and substructures. This load can be counteracted by balancing masses at the crankshaft webs. The configuration of this mass balancing depends on the number and arrangement of the cylinders as well as the distribution of the crank throw.

Inertial Force Balancing on Single-Cylinder Machines. Only rotating and oscillating inertial forces, but no moments of inertia, occur on a single-cylinder machine since the plane of symmetry lies in the cylinder axis.

The rotating inertial forces can be balanced relatively easily and completely by two counterweights m_{rG} on the crankshaft (Fig. 10.16). These are offset by 180° with respect to the crankshaft journal and are calculated from the rotating mass m_r and the corresponding distances from the center of rotation as

$$m_{rG} = 0.5 m_r \frac{r}{r_G}\,, \tag{10.45}$$

where r is the crank radius and r_G is the distance of the counterweights from the center of rotation.

The distribution of the balancing masses to individual counterweights with the mass m_{rG} is incorporated to prevent an additional moment.

To balance the oscillating inertial forces, (10.38) is applied and broken down into first- and second-order forces of inertia. The change of the first-order forces of inertia corresponds to the frequency of the crankshaft speed and consequently can be partially balanced by counterweights on the crankshaft.

Second-order forces of inertia change twice as fast. Therefore they cannot be balanced by counterweights on the crankshaft.

The oscillating inertial forces act in the direction of the cylinder axis. Hence, only the perpendicular force components of the rotating counterweights can be used as mass balancing. The horizontal force components also generated represent an interfering force. The following equation is used to calculate the counterweights

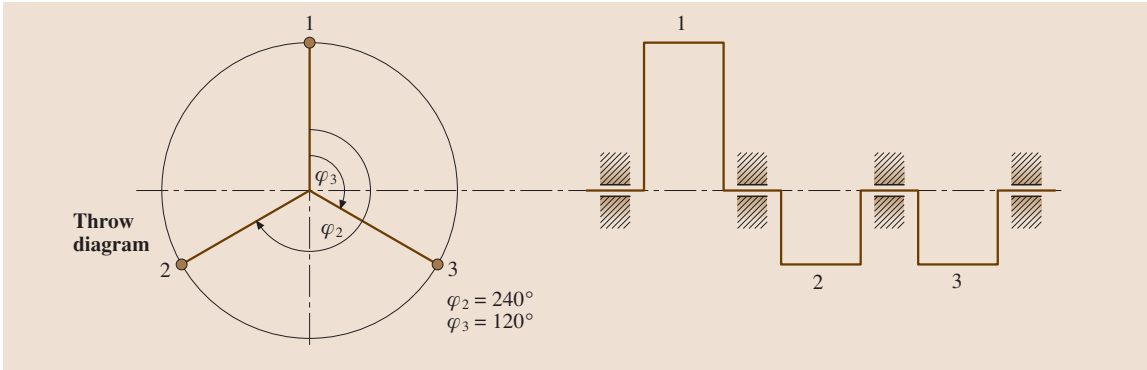


Fig. 10.17 Cross and longitudinal sections of a three-cylinder crankshaft (after [10.3])

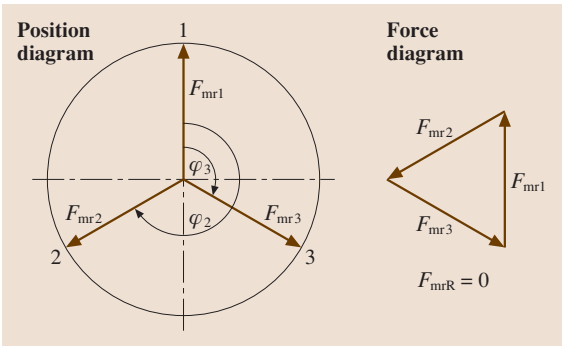


Fig. 10.18 Determination of the resultants from the rotating inertial forces (after [10.3])

m_{oIG} per web to balance the first-order forces of inertia

$$m_{oIG} = 0.5 m_o \varphi \frac{r}{r_G} , \qquad (10.46)$$

where φ represents the proportion of the first-order forces of inertia being balanced. This prevents the influence of the interfering horizontal force components from becoming too great.

Mass Balancing on Multi-Cylinder Machines. To balance masses for multi-cylinder engines, the inertial forces are calculated for every cylinder and consolidated in a resultant. The aim of mass balancing is to compensate the inertial forces reciprocally and to keep the resultants as small as possible.

Furthermore, the concentrated loads not engaging in the center of mass cause additional moments of inertia. These are also consolidated into resultants.

The radial and axial arrangement of the crank throw strongly influence the value of the resultants. When designing the crankshaft, attention has to be paid to balancing the concentrated loads. One simple way to do this is by projecting the crankshaft radially and axially. Cross and longitudinal sections of the crankshaft are drawn (Fig. 10.17). The cross section (the left-hand image in Fig. 10.17) is called a throw diagram because of its shape. Along the longitudinal section (the right-hand image in Fig. 10.17) of the crankshaft, the crank throws are numbered from left to right and subsequently transferred to the throw diagram.

Resultants from the Rotating Inertial Forces. The rotating inertial forces calculated with (10.32) are transferred in parallel to a common cross-sectional plane. By adding the vectors, the individual forces are consolidated into a resultant F_{mrR} (Fig. 10.18).

The resultant rotates with the crankshaft speed and has a constant value. Hence it does not have to be determined anew for other crankshaft positions, but only rotated by the corresponding angle (Fig. 10.18).

Resultants from the First-Order Forces of Inertia. The specific instantaneous value in the cylinder axis is calculated by projection onto these and the total first-order inertial forces in the direction of the cylinder axis is determined by adding their vectors. The maximum value of the first-order forces of inertia is calculated with

$$F_{moI\_max} = m_o \omega^2 r . \qquad (10.47)$$

A simpler approach is often used in practice. The maximum values are likewise applied in the direction of the crank arm. Adding these forces yields a resultant of the maximum values $F_{moI\_max\,R}$ that is projected onto the cylinder axis. The force calculated in this way is the resultant of the first-order forces of inertia F_{moIR} (Fig. 10.19).

The resultant of the maximum values rotates together with the crankshaft as it rotates. In the corresponding crank position, it must be projected

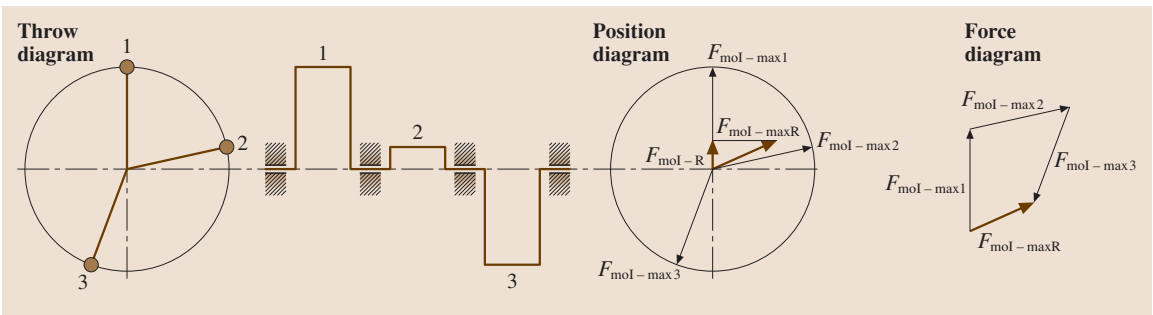


Fig. 10.19 Determination of the resultants from the first-order forces of inertia (after [10.6])

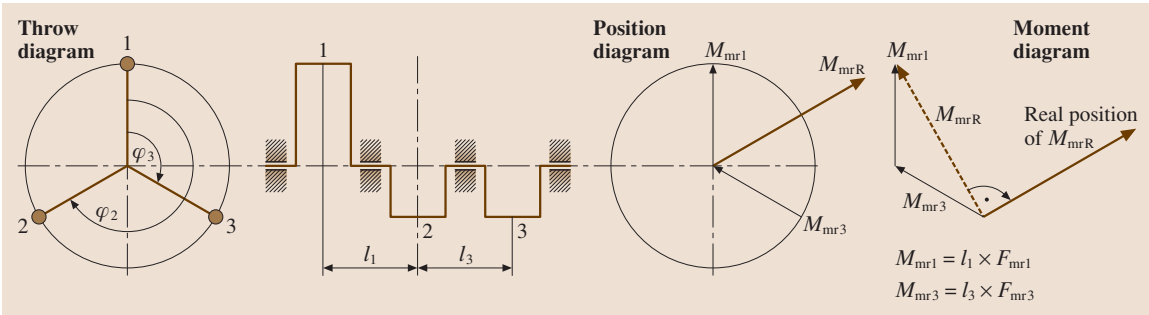


Fig. 10.20 Determination of the resultant moment of inertia from the rotating inertial forces (after [10.6])

anew onto the cylinder axis in order to obtain the corresponding resultant of the first-order forces of inertia.

Resultants from the Second-Order Forces of Inertia. The resultant of the second-order forces is calculated in the same way as the resultant from the first-order forces of inertia. However, the second-order forces of inertia change their value as the crankshaft's angle of rotation doubles. Therefore they cannot be applied in the direction of the crank arms. Rather, they have to be drawn at twice the angle. A force plan is again used to calculate the resultant of the maximum values, which is subsequently transferred to the position diagram. Projection onto the cylinder axis yields the resultant of the second-order forces of inertia.

The rotation of the resultant of the maximum values has to be entered on the site plan at twice the crankshaft angle. The resultant sought is calculated by projection onto the cylinder axis.

Resultant Moment of Inertia. The inertial forces in multi-cylinder engines act at a specific distance to the machine's center of gravity and cause moments of inertia. Since their precise determination is extremely involved, a simplified assumption places the center of gravity in the longitudinal section in the crankshaft axis (e.g., the second cylinder's axis in a three-cylinder machine; see the longitudinal crankshaft section in Fig. 10.20). The resulting error is negligible in most cases.

The moment vectors are applied corresponding to the crank arms taken from the throw diagram in the site plan. The direction of the vectors is determined as follows. The vectors in the longitudinal section to the left of the reference point point outward, while the vectors to the right of the reference point point toward the center of the throw diagram.

The value of the resultant moment of inertia is obtained by adding the vectors in the moment diagram. The vector is given its correct position by rotating it clockwise by 90°.

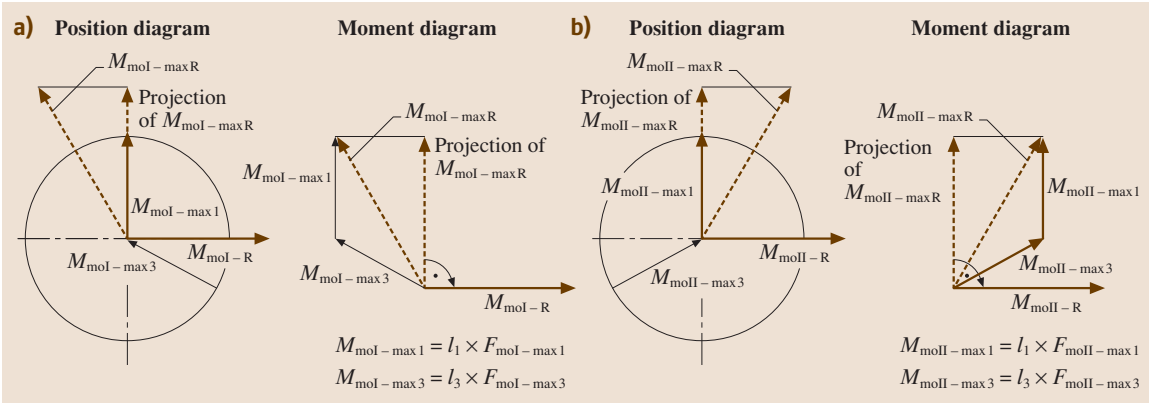


Fig. 10.21a,b Determination of the resultant moment of inertia from the oscillating **(a)** first- and **(b)** second-order inertial forces (after [10.6])

A similar approach is employed to determine the resultant first-order moment of inertia. However, the resultant of the maximum values of the moments is projected onto the cylinder axis and then rotated clockwise by 90° (left-hand position and moment diagram in Fig. 10.21).

When the resultant second-order moment of inertia is calculated, the moment vectors must be entered at twice the crank angle (see the right-hand position and moment diagram in Fig. 10.21). Just as with the resultant first-order moment of inertia, the resultant of the maximum values is projected onto the cylinder axis and rotated by 90°.

10.1.4 Selected Elements of Reciprocating Machines

Crankshaft Drive

The crankshaft drive can be constructed in a plunger piston or crosshead design. The crosshead relieves the piston from lateral piston force, especially in large piston machines (Fig. 10.22).

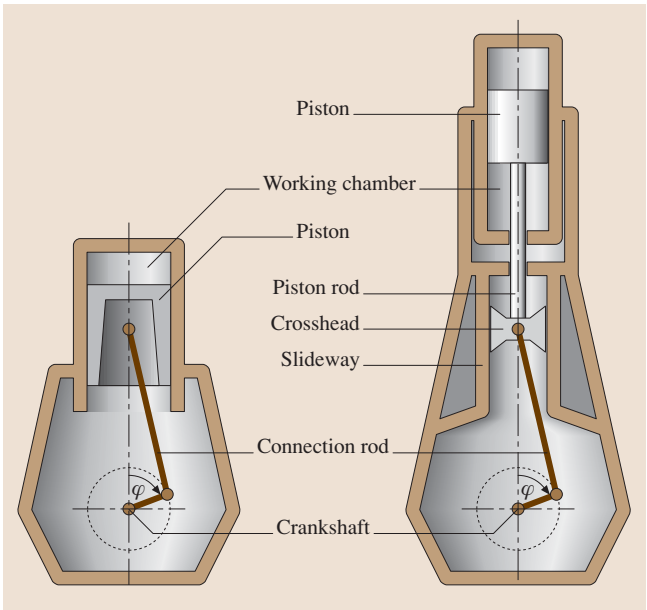


Fig. 10.22 Plunger piston and crosshead crankshaft drive (after [10.6])

Crankshaft

The crankshaft shown in Fig. 10.23 has a shaft journal (1) running in bearings connected by webs (3) to the crankshaft journal. Counterweights (4) on the webs balance the rotating inertial forces. As a rule, the crankshaft is suspended by $z+1$ main bearings, where z is the number of cylinders.

The crankshaft is under stress from forces and bending and torsional moments. The webs are dimensioned accordingly to handle the high stresses.

Different methods are used to manufacture crankshaft blanks. A difference is principally made between cast and forged crankshafts. Cast crankshafts weigh 10% less than forged ones because of the low density of the nodular graphite cast iron frequently used. Giving cast crankshafts a hollow design can boost this even more. A significant disadvantage of cast over forged crankshafts is the low elastic modulus of penetration and associated lower stiffness. In Europe, the market share of cast-iron car and truck crankshafts that are not highly stressed is 60% [10.7].

The most important forging methods can be classified as open die forging and closed die forging. While hammer forging is only used for prototypes and custom-made pieces, closed die forged crankshafts are primarily used in cars and trucks in conjunction with large lot sizes.

If the size of the crankshaft being produced exceeds the production possibilities of forging and casting, the shaft and crankshaft journal are directly connected to one another by the webs. These so-called multi-piece crankshafts are primarily used in the manufacture of large diesel engines.

Forged and cast crankshafts are subjected to different post-processing in order to increase a crankshaft's component strength. Apart from inductive hardening to increase the bearing journal's resistance to wear, rolling the radii in the transition bearing journal and web and nitriding build up residual compressive stresses in the journal and radii zone. As a result, the fatigue strength can be increased considerably in these highly stressed zones.

Connecting Rod

The connecting rod transmits forces between the piston and crankshaft journal. It is not only under stress from tension and compressive forces but is also subjected to bending by the respective inertial forces. This necessitates a design that is both rigid and light, especially in high-speed combustion engines.

The connecting rod (Fig. 10.24) consists of a shank and two conrod eyes, which act to connect the piston and the crankshaft. It is cast, die forged or sintered. Heat-treated steel as well as gray cast iron, malleable iron and light metals are used as materials. Fracture separations, also called cracks, separate (straight or

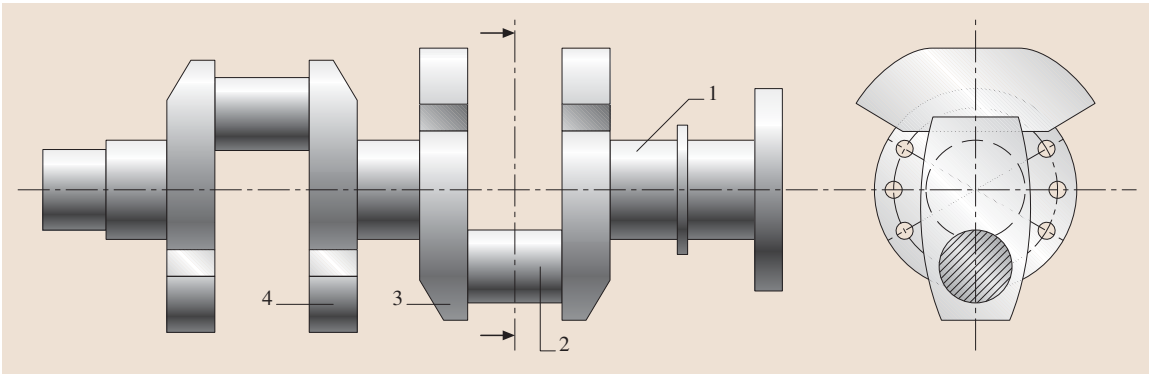


Fig. 10.23 Elements of a crankshaft (after [10.4])

obliquely) the bottom (large) small-end bearing. Anti-fatigue bolts hold the halves of the large conrod eye together.

Piston

The piston's job is to transmit fluid forces to the connecting rod or connecting rod forces to the fluid (Fig. 10.25).

In addition, the piston seals the working chamber by means of piston rings. The number of piston rings is dependent on the pressure difference between the upper surface and the bottom surface. In combustion engines, for example, the piston rings dissipate the heat generated in the pistons, which piston cooling can subsequently pass to the motor oil.

Hence, having high resistance to wear and resistance to heat as well being able to conduct heat and being lightweight are the most important requirements on a piston.

This is why pistons are usually made of light alloys and less often gray cast iron (smaller pistons). Low piston weight has a positive effect on the oscillating inertial forces. Steel and cast steel are only used for the top part of multi-piece pistons or for plunger pistons.

Lubrication

The lubrication of piston machines serves the functions:

- Forming a stable lubricating film between the structural parts
- Reducing friction
- Reducing wear in the machine
- Cooling and cleaning bearings and sliding surfaces
- Forming a seal between the piston ring and cylinder wall

Pressure circulation lubrication is primarily used. The oil is delivered by a pump that is powered by a piston machine or electrically powered. It is purified, cooled and subsequently pumped into the main oil gallery. From there, it reaches the lubricating points such as the crankshaft, valve gear and piston.

Cooling

The heat losses occurring on the walls of the working chamber necessitate cooling of the corresponding components. The heat flow dissipated is dependent on the surface, the heat transfer coefficients and the temperature difference. Water and air are mainly used as coolants and prevent overheating of the components and lubricant as well as power losses caused by filling losses. However, water's better cooling effect increases the complexity of designing as well as manufacturing of the machine. Water-cooled machines are given a cooling jacket through which water is pumped. In the case of air cooling, the design of

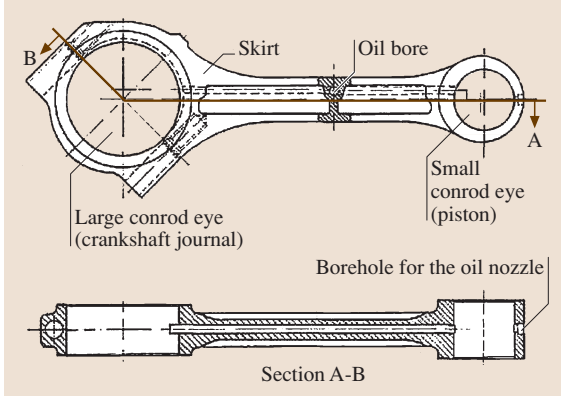


Fig. 10.24 Connecting rod design (after [10.6])

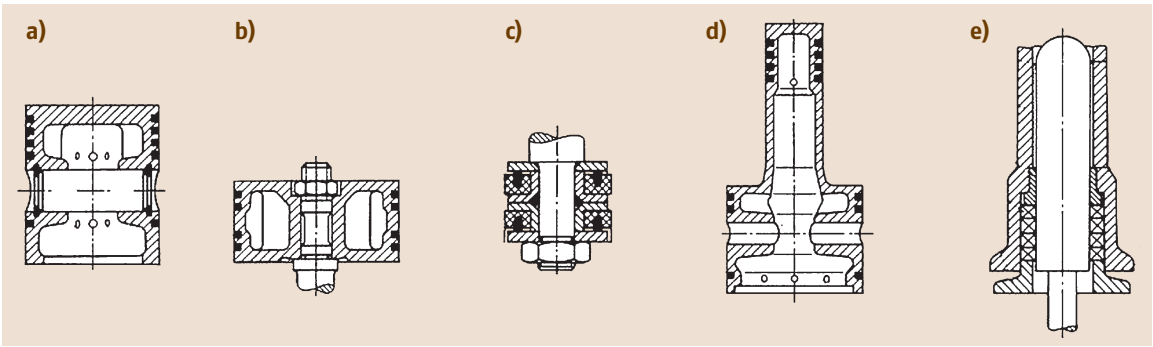


Fig. 10.25a–e Piston variants (after [10.4]). **(a)** Plunger, **(b)**, **(c)** disc pistons, **(d)** stepped piston, **(e)** plunger piston

the fins enlarges the surface of the components being cooled.

Suspension

As a rule, a split bearing is used for the suspension in reciprocating machines because of the offset shape of the crankshaft. Exceptions are multi-piece crankshafts or small and medium-sized piston compressors in which antifriction bearings, among others, can be used.

Their simple design and the small space required in the crankcase, their ability to absorb impacts and vibrations as well as their low mass is increasingly leading to friction bearings with hydrodynamic lubrication being used in reciprocating machines.

Apart from great strength, the materials used must also have optimal tribological properties. Hence, friction bearings are usually designed with a harder matrix (e.g., CuSn, AlCu) into which a soft, low-melting-point material (e.g., Pb, Sn) is incorporated.

The suspension's reliability is crucially important for guaranteeing a piston machine's operation. Hence, optimal engineering utilizes improved bearing materials and state-of-the-art calculation methods.

10.2 Positive Displacement Pumps

10.2.1 Types and Applications

A positive displacement pump is a work machine for incompressible media (fluids), the working chamber of which is periodically altered by the displacer (piston). The displacer's motion can be either oscillating (reciprocating pump) or rotating (rotary piston pump).

Reciprocating pumps mainly use automatically operating, pressure-controlled valves to control the process, i. e., to connect and disconnect the working chamber to the suction and the pressure line. Rotary piston pumps use ports to do this.

In each case, the discharge of the medium directly follows the characteristic of the displacer's motion (volume displacement).

Theoretically, the maximum pressure achievable with positive displacement pumps is unlimited and determined only by internal leakage losses, the medium's compressibility, the component strength and the motive power.

During delivery, mechanical energy is transferred to the medium as potential energy, which serves to compensate for level or pressure differences.

A difference is made between reciprocating pumps with rigid pistons (disc and plunger pistons and plungers, Fig. 10.26) and with elastic pistons (diaphragm, hose, Fig. 10.27).

Radial and axial piston pumps in which the piston is typically designed as a plunger are also used in high-pressure hydraulics. Depending on the design (radial or axial), the stroke motion in these pumps is generated by a rotating eccentric cam or a swash plate and a fixed cylinder unit, i. e., similar to reciprocating pumps with crankshaft drive.

In part, the lift stroke for a fixed eccentric or swash plate is also produced with a rotating cylinder unit (drum/star). In systems with a rotating cylinder unit, control blocks with ports are used instead of the working valves otherwise customary in reciprocating pumps (Figs. 10.28, 10.29).

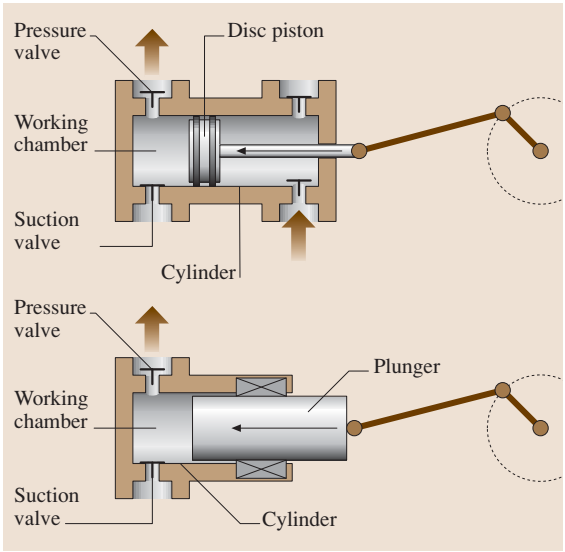


Fig. 10.26 Rigid piston design

Rotary piston pumps are classified according to the primary motion of the displacer as single rotation piston machines, planetary rotation piston machines and circulation piston machines (Figs. 10.30–10.33).

Other designs for delivering media with solid fractions or for liquefied gases or even with direct drive without crankshaft drive, e.g., steam and pneumatic pumps or hand pumps, are characterized in every configuration by special requirements and have different, usually oscillating types of motion, which for the most part are directly generated without a crankshaft drive or the like (Fig. 10.34).

Furthermore, other designs exist that cannot be directly classified in the categories cited because of their types of motion. One example is the peristaltic pump used in medicine and laboratories (Fig. 10.35).

The vane pump, primarily used for garden irrigation and in small boats, is simple in design and durable in operation (Fig. 10.36).

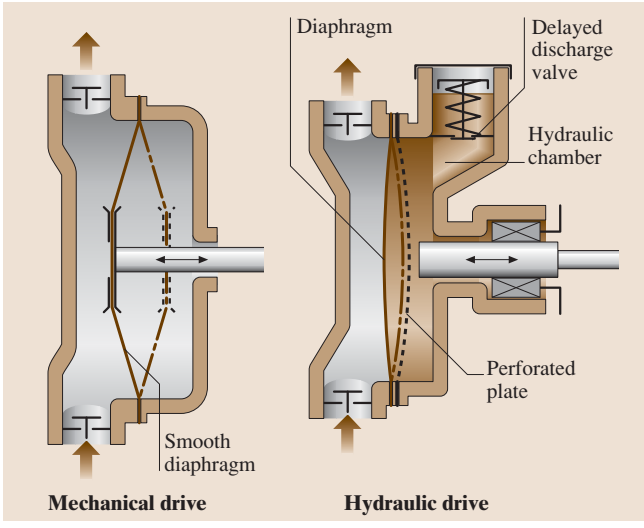


Fig. 10.27 Diaphragm designs

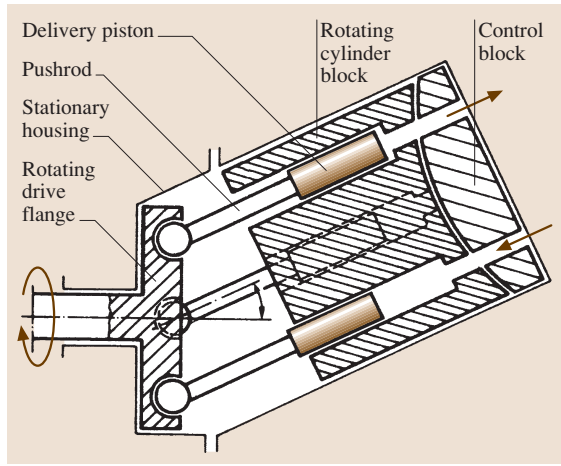


Fig. 10.28 Axial piston pump

10.2.2 Basic Design Parameters

Figure 10.37 presents the principle of reciprocating pump operation.

The crankshaft drive, consisting of the crankshaft, connecting rod and crosshead, converts rotary drive mo-

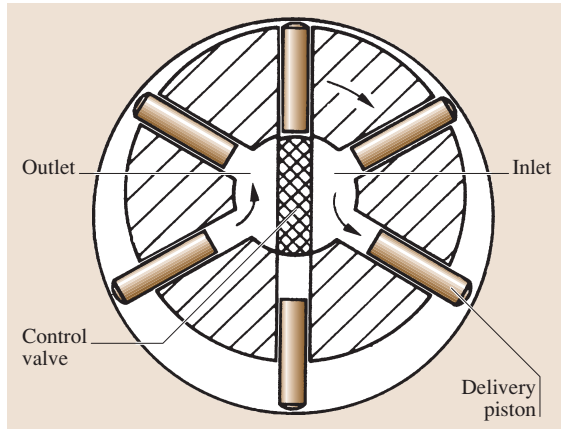


Fig. 10.29 Radial piston pump

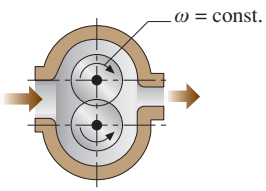
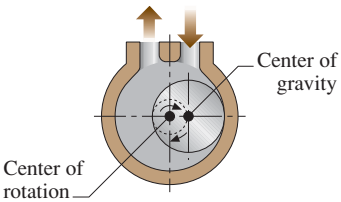
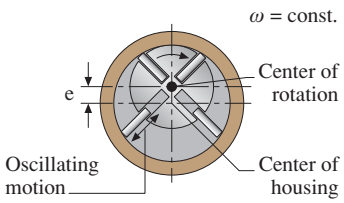
Single rotation piston machines	Planetary rotational piston machine	Circulation piston machine
Moving parts induce uniform rotation around their centers of mass.	The center of gravity of the uniformly rotating displace describes a circular path.	Along with uniformly rotating parts there are also irregularly moving parts that execute circular or circle-like or even oscillating motions.
Moving parts are directly balanced: bearings and shafts are not subjected to stress from centrifugal forces → High speeds are possible.	The rotary imbalance can only be directly balanced. Increased stresses on bearings caused by centrifugal forces. Speeds are limited.	Increased friction in the machine → overall efficiency decreases. Wear increases.
Examples: Gear pump, screw design, roots blower, gear tooth design, liquid ring design.	Examples: Trochoidal-type design, eccentric screw pump, eccentric design, rolling piston design.	Examples: Multi-cell or impeller cell design, rotary valve design, external-vane design.
Operating principle  (Gear pump)	Operating principle  (Eccenter-type design)	Operating principle  (Vane-type design)

Fig. 10.30 Rotary piston machines

tion into an oscillating motion, which the piston rod transmits to the piston. In the cylinder, the piston executes a motion between the two dead center positions, the crank dead center (CDC) and the head dead center (HDC), whereby the volume trapped between piston and cylinder cover diminishes and expands periodically.

In principle, the function and process are identical for rotary positive displacement pumps. The resultant working cycle can be broken down as follows:

1. Induction. The volume of the working chamber expands during the piston's motion from CDC to HDC. This produces a vacuum in the working chamber causing the suction valve (SV) to open. This vacuum causes the fluid delivered from the suction line to flow into the cylinder. At nearly constant pressure, induction theoretically occurs during the complete piston stroke (corresponding to line 1–2 in the p–V diagram).
2. Pressure rise. The piston motion reverses in the crank dead center. The resultant reduction in volume causes the pressure in the working chamber to rise and the suction valve to consequently close. Since liquids are theoretically incompressible, the pressure rises at a constant volume to the pressure at the pressure valve (PV) (corresponding to line 2–3 in the p–V diagram).
3. Expulsion. The reduction in volume during the piston's motion from CDC to HDC causes the pressure in the working chamber to increase. When a pressure is attained, which dominates on the pressure side of the pump, the pressure valve opens. The fluid is expelled approaching the pressure, which dominates at the outlet (corresponding to line 3–4 in the p–V diagram).
4. Pressure drop. When the head dead center is reached, the piston motion reverses again so that the pressure drops at constant volume causing the pressure valve to close and the suction valve to open (corresponding to line 4–1 in the p–V diagram).

Whereas the crankshaft has a constant angular velocity (speed), the piston speed in the cylinder is not constant. It is zero at the dead centers. During its motion from one dead center to the other, there is acceleration and deceleration, each corresponding to the crankshaft's law of motion (Sect. 10.1.3). The piston speed is a function of the angle of rotation α and, as a function of the crankshaft drive's connecting rod ratio (10.23), has an approximately sinusoidal characteristic line.

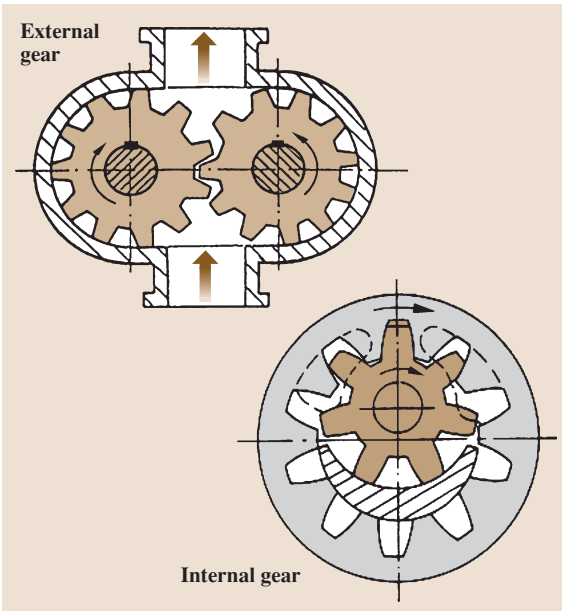


Fig. 10.31 Single rotation piston design

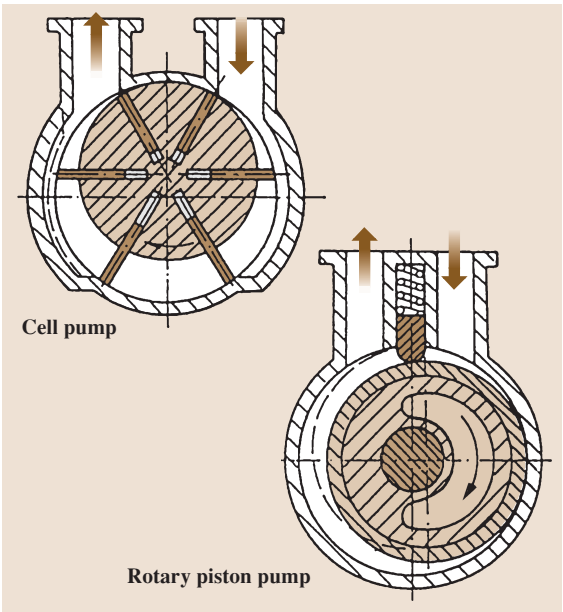


Fig. 10.33 Circulation piston design

Operating Behavior

The p–V diagram in Fig. 10.37 represents an ideal development of pressure, presuming a massless fluid flows without loss. Since losses and inertial forces always have an effect under real conditions, the real p–V diagram of a piston pump in Fig. 10.38 is produced. The ideal pressure development from Fig. 10.37 has been incorporated for comparison.

- Since the frictional losses in the suction line and the suction valve have to be compensated for, suction occurs at a pressure lower than the suction tank pressure. Moreover, another drop in pressure is needed to open the suction valve since the valve opening resistance (valve acceleration, surface ratios at the valve seat, valve elastic force) has to be overcome. The mass inertia of the decelerated liquid column or the suction pressure can also cause a slight rise in pressure toward the end of the intake stroke.
- Since the mass inertia does not cause the suction valve to close abruptly when the CDC is reached and a real fluid is not fully compressible, the pressure does not rise at a constant volume. Moreover, fluids frequently contain small quantities of gas that cannot be compressed. As a result, the real pressure rise from 2′ to 3′ is not isochoric.
- Expulsion occurs at a higher pressure than on the pressure side of the pump since the corresponding losses again have an effect here. Analogous to the

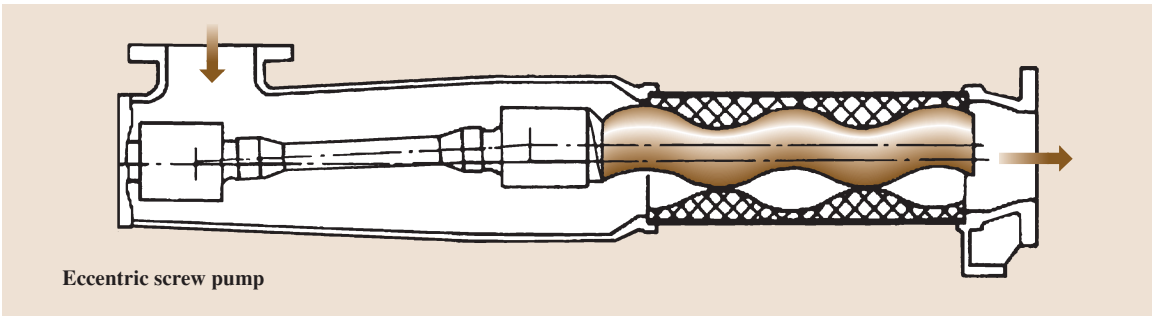


Fig. 10.32 Planetary rotation piston design

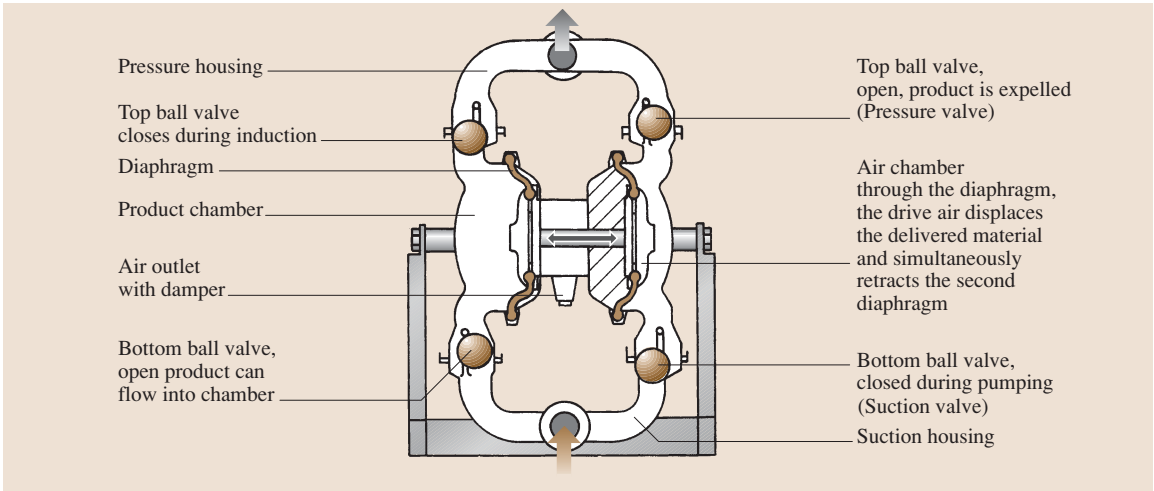


Fig. 10.34 Compressed-air diaphragm pump (after [10.8])

suction valve, an additional increase in pressure is needed to open the pressure valve.

- In turn, the pressure cannot drop at a constant volume since there are inertial forces and potential back-flows from the pressure line when the pressure valve closes and the gas fractions re-expand. As a result, the ideally isochoric pressure drop from $4'$ to $1'$ is not isochoric.

In contrast to centrifugal pumps, piston pumps have a characteristic curve with a delivery rate that is nearly independent of delivery pressure. As a result of the displacement, the delivery rate is calculated as the product of the piston surface s, stroke and rotational speed n

$$\dot{v} = Q = A_K s n \,. \tag{10.48}$$

Consequently, throttling, i. e., changing the system head curve, cannot regulate the delivery rate. On the other hand, an inordinate rise in pressure has to be prevented from destroying the pump and system components.

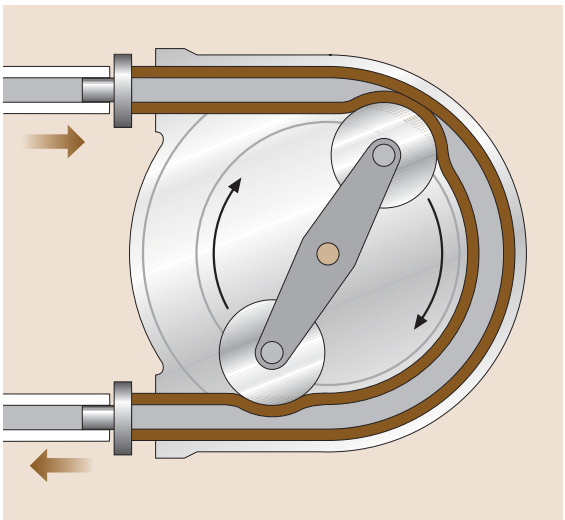


Fig. 10.35 Peristaltic pump (after [10.9])

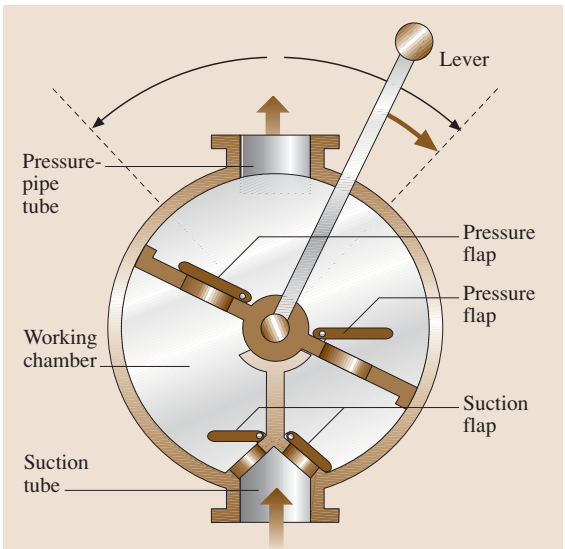


Fig. 10.36 Sliding vane pump

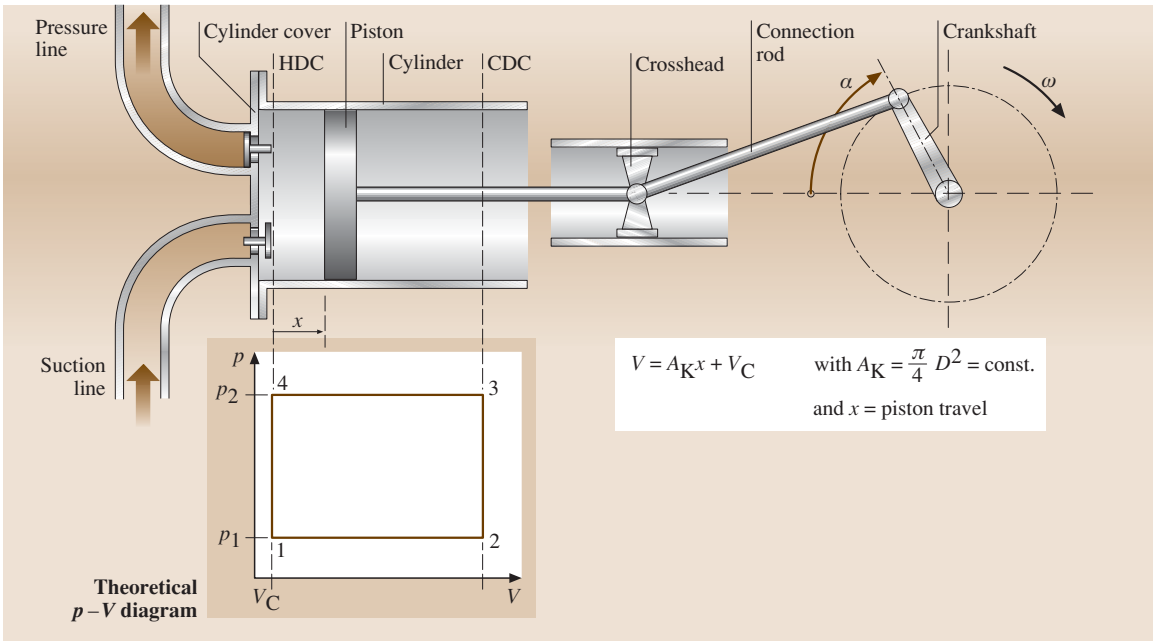


Fig. 10.37 Principle of reciprocating pump operation

Hence, any time a piston pump is used, a *safety valve* must be incorporated on the system's pressure side.

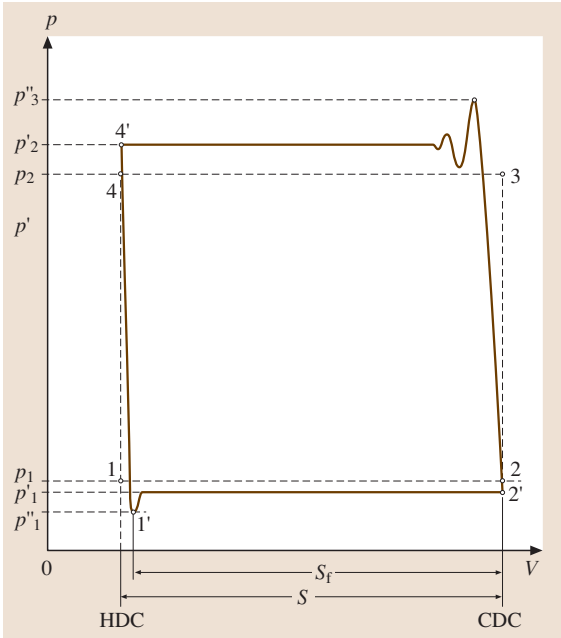


Fig. 10.38 Ideal and real piston pump p–V diagram

Energy and Efficiency

Specific Effective Energy of the Complete System w_{st}. Figure 10.39 shows a simple pump system.

The pump feeds mechanical work to the fluid in order to increase the pressure of the potential energy, i. e., to facilitate delivery and/or a rise in pressure.

If this work is applied to the fluid's mass (disregarding changes in density), then the specific effective energy of the system is obtained

$$w_{st} = w_{geo} + \frac{(p_a - p_e)}{\rho} , \qquad (10.49)$$

$$w_{geo} = H_{geo} g , \qquad (10.50)$$

where w_{geo} is the specific geodetic delivery work, p_e is the pressure on the fluid in the suction tank, p_a is the pressure on the fluid in the pressure tank, ρ is the density of the pumped medium, g is the gravitational acceleration, and H_{geo} is the level difference.

Expressed in pressures and levels, the following ensues

$$p_{st} = H_{geo}\rho g + p_a - p_e , \qquad (10.51)$$

$$H_{st} = H_{geo} + \frac{(p_a - p_e)}{\rho g} . \qquad (10.52)$$

The system's specific effective energy is independent of the type of pump, the size of the delivery flow, and the layout of the pipe system.

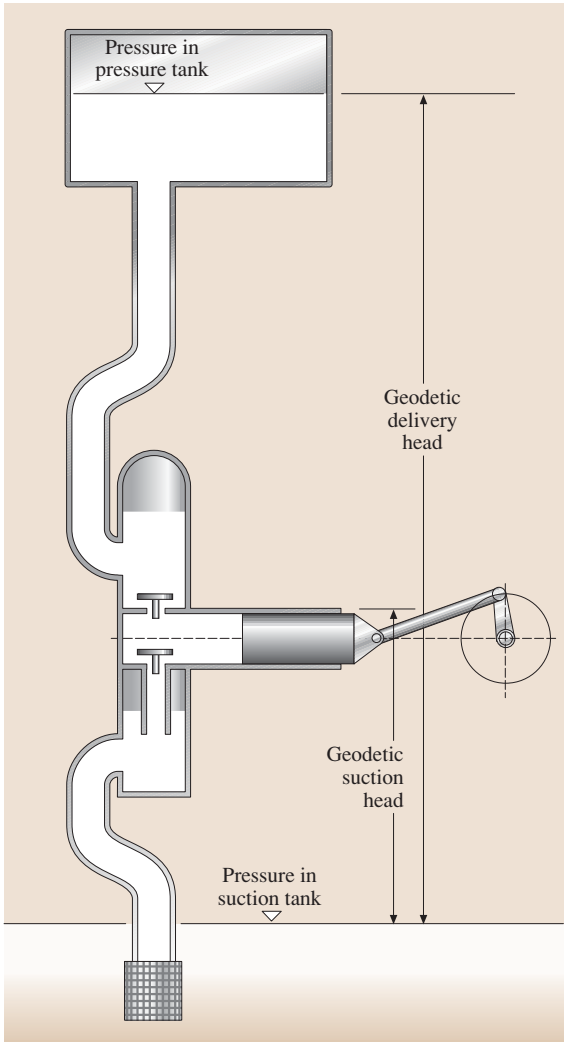


Fig. 10.39 Piston pump system

Specific Effective Energy of the Pump w. The specific delivery work required to deliver the mass flow of the fluid in the selected system is referred to as the specific effective energy of the pump. The proportions of energy loss and velocity energy are taken into account here

$$w = w_{st} + w_{dyn} + w_V \,, \tag{10.53}$$

$$w_{dyn} = 0{,}5\left(v_2^2 - v_1^2\right) \,, \tag{10.54}$$

where v_2 and v_1 are the velocities in the suction and pressure tubes, respectively, and w_v is the specific energy loss in the suction and pressure lines.

The pump's specific effective energy depends on:

1. The type of pump
2. The size of the delivery flow and
3. The layout of the pipe system

Delivery Rate η_{Vol}. As a result of volume losses, a pump's effective delivery rate is lower than the theoretical rate. These losses are caused by:

1. Leaks or valve-closing delays (volumetric efficiency λ_L)
2. The presence of gases or incomplete filling (λ_F)

The product of λ_L and λ_F constitutes the delivery rate η_{Vol}. The following empirical values apply to the delivery rate η_{Vol}:

1. 0.88–0.92 for small pumps
2. 0.92–0.96 for medium-sized pumps
3. 0.96–0.98 for large pumps

Indicated Efficiency η_i. Indicated efficiency expresses the ratio of a system's effective power to a pump's indicated power (determined from the indicator diagram). Thus, hydraulic losses are also taken into account.

$$\eta_i = \frac{P_Q}{P_i} \,, \tag{10.55}$$

$$P_Q = \rho Q w_{st} \,, \tag{10.56}$$

where Q is the feed rate and P_i is the indicated power (from the indicator diagram).

Mechanical Efficiency η_m. Mechanical efficiency expresses all the losses caused by *mechanical friction* inside the pump. In the process, the ratio of indicated power to coupling power (input power) is expressed by

$$\eta_m = \frac{P_i}{P_{zu}} \,. \tag{10.57}$$

Experience has shown that pumps with crankshaft drive have a mechanical efficiency in the range

$$\eta_m = 0.85 - 0.95 \,. \tag{10.58}$$

Total Efficiency of the Pump η. Total efficiency η is expressed by the ratio of the effective power to the input power

$$\eta = \frac{P_Q}{P_{zu}} \,. \tag{10.59}$$

Experience has shown that a pump with crankshaft drive has a total efficiency in the range of $\eta =$ 0.70–0.80.

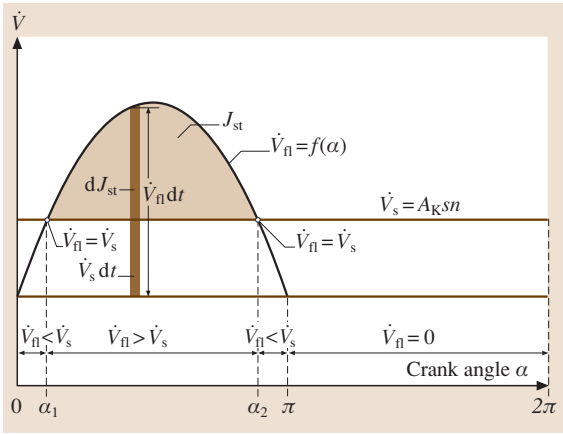


Fig. 10.40 Time progression of the volumetric delivery of a single-action one cylinder pump

Mass Actions

On the one hand, the reciprocating pump's discontinuous working process results in a constantly fluctuating speed of the pumped medium in the connected lines. On the other hand, the portions of the pumped medium subjected to accelerations undergo a mass action. The time-varying delivery rate follows the laws of piston motion up to a fixed limit (the vapor pressure) so that the time progression of the volumetric delivery corresponds exactly to the piston velocity. The valves' action is used to allocate the positive and negative fractions to the intake and the delivery stroke (Fig. 10.40).

The crankshaft's rotation is split into the intake and the delivery stroke. Thus the total delivery rate must be pumped during a half rotation of the crankshaft. The maximum value of the speed is considerably larger than the average value. There are only two times during both the intake and the delivery stroke at which the instantaneous delivery rate corresponds exactly to the average volumetric flow. This is not the case at any other time.

The maximum acceleration is at a crank angle of 0°, i. e., the theoretical start of the intake stroke. However, a maximum acceleration of the amount of fluid also corresponds to a maximum mass action, which results in a maximum reduction in the pressure in the working chamber (Fig. 10.41).

To prevent cavitation, the so-called vapor pressure p_0 of the fluid must not be reached during pump operation. Consequently, the piston acceleration and the mass of the fluid present in the intake tract yield the

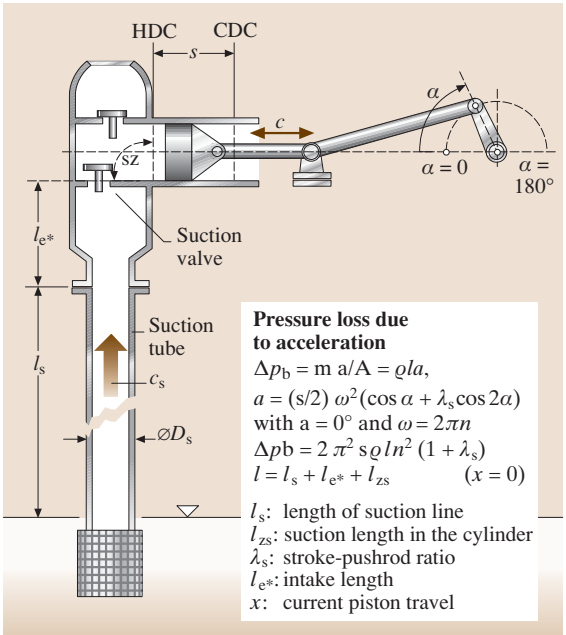


Fig. 10.41 Suction side of a reciprocating pump system

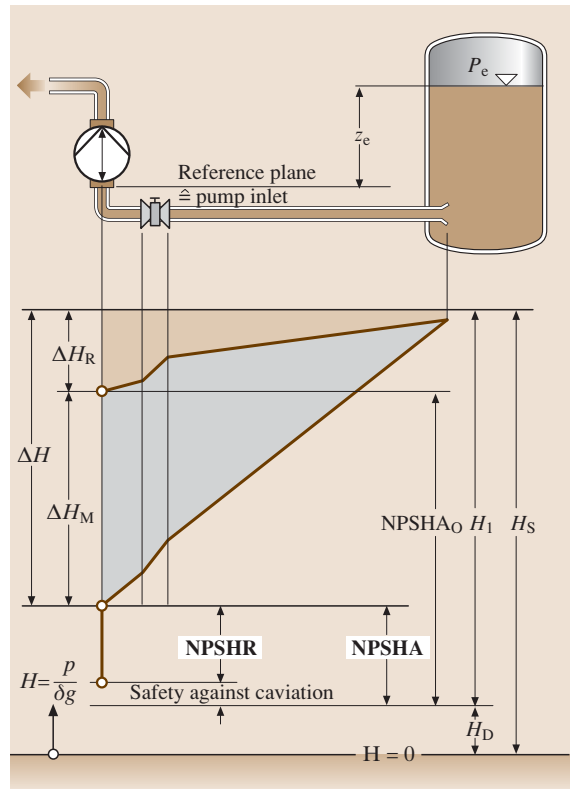


Fig. 10.42 NPSH value of the intake tract of a pump system (after [10.10])

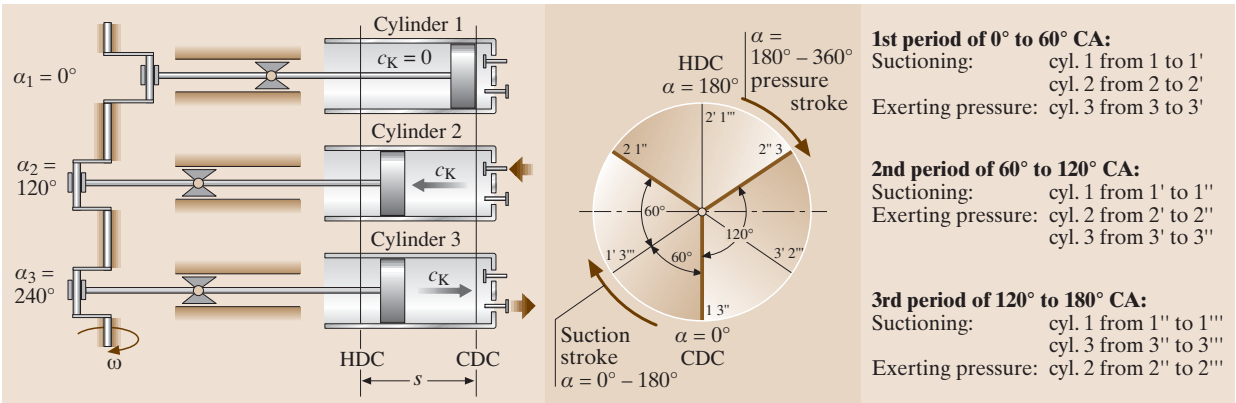


Fig. 10.43 Function of a three-cylinder pump

maximum permissible suction head. Conversely, when the suction head is known, it yields the maximum permissible speed for the pump.

The net positive suction head (NPSH; Fig. 10.42) is used to designate a pump's volume flow rate. The difference between the actual and the required NPSH is a measure of protection against cavitation.

When several working cylinders are employed, conditions change with the effect that several cylinders are able to draw from one suction line at the same time (Fig. 10.43). Here, the drop in pressure as a result of the mass action must be determined individually for the single parts of the intake tract. Employing several cylinders improves the overall balance since the entire amount of fluid is not subjected to maximum acceleration and, when there are upwards of three working chambers, the flow velocity in the suction line is never zero.

To reduce the pulsation of the delivery rate both on the suction and the pressure side, employing several cylinders is expedient.

If using several cylinders proves insufficient, then installing elastic components, the so-called expansion chamber, is advisable. Here, interconnecting an elastic accumulator (gas cushion) achieves a homogenization of the delivery rate.

The cyclic irregularity of the the pressure on the piston is consulted as a decision criterion for integrating an expansion chamber

$$\delta_p = \frac{(p_{K-max} - p_{K-min})}{p_{K-m}}, \qquad (10.60)$$

where

$$p_{K-m} = \frac{p_{K-max} + p_{K-min}}{2}. \qquad (10.61)$$

This applies to both the suction and the pressure side.

The cyclic irregularities considered limit values for the use of expansion chambers are $\delta_{p-s} \leq 0.1-0.05$ (suction side), and $\delta_{p-d} \leq 0.05-0.02$ (pressure side).

The expansion chambers can be constructed both as a simple air tank open toward the fluid (Fig. 10.44) or as a diaphragm or bladder accumulator (Fig. 10.45).

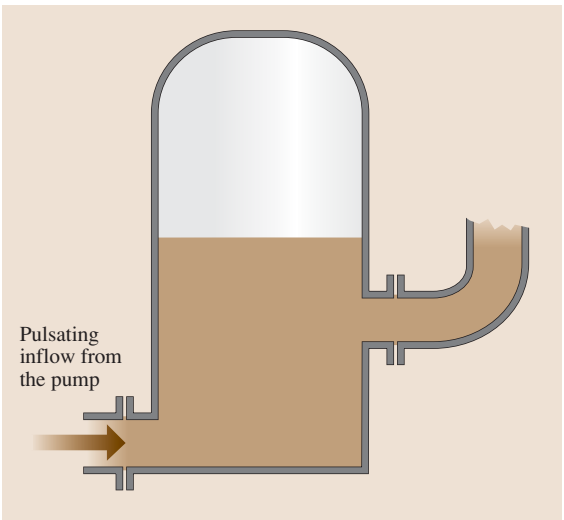


Fig. 10.44 Pressurized air chamber

10.2.3 Components and Construction of Positive Displacement Pumps

Working Valves

Automatically operating, pressure controlled valves are used in reciprocating pumps to separate the suction and expulsion processes. These valves can be constructed as

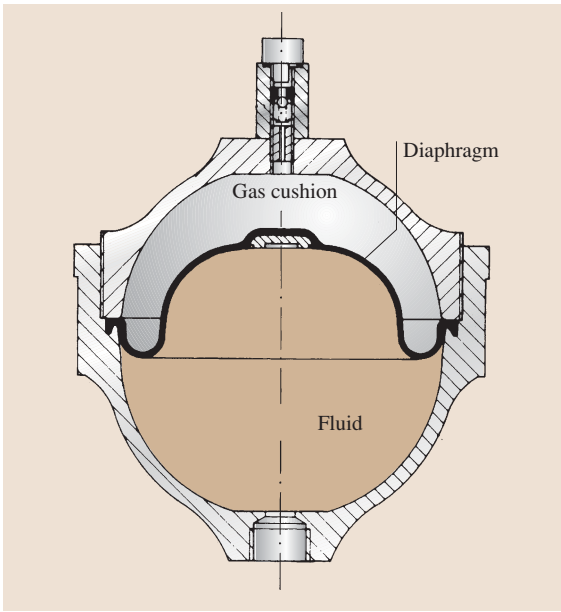


Fig. 10.45 Diaphragm accumulator (after [10.11])

flat seat, conical seat or ball valves. Both, elastic force and gravitational force can trigger their closing forces (Fig. 10.46).

Multiple annular valves (Fig. 10.47) are also used to reduce the force of acceleration at the valve and above all to lower their speed when they close. These make large openings possible when strokes are small.

To open such a valve, a pressure differential must always be present, which essentially depends on the elastic force, the friction and the valve face (Fig. 10.48).

Reciprocating Pumps

Compression Pumps. Compression pumps are pumps that, as boiler feed pumps for example, work approaching pressures up to 7000 bar and achieve delivery flows of 0.01–160 m^3/h. As a result, pump components are under most extreme stresses, most notably in the cylinder (called the pump head here) the valves and the plunger seal (stuffing box). In view of the high pulsating stress of the components, the design not only has to apply appropriate methods of calculation and employ the requisite material but also produce particular surface finishes to ensure sufficient operational reliability.

For economic reasons, such pumps, which are usually only manufactured in small quantities, are often engineered based on *standardized* engines with crossheads. Thus, they are then designed for specific forces along the rod (piston forces). When standard engines are used for different pumps, a specific maximum piston diameter is produced for every pressure to be reached. The number of cylinders and the speed determine the delivery flow pumped. By combining pump heads in these engines, appropriate pumps can be assembled (modular system) for the widest variety of applications with relatively little effort.

The design engineering and material selection for individual components such as the plunger, pump cylinder, valves, etc. are extremely important for pumps with higher pressures. Composite materials or prestressed components are used at extreme pressures (Fig. 10.49).

An important aspect of design is making the disassembly of wearing parts (valves, stuffing box packing) as easy as possible. Above all, it is essential to avoid having to disassemble the suction and pressure lines when removing the valve since these are very rigid at extreme pressures.

The stuffing box packings in compression pumps are under extreme stress and therefore may not be used as a plunger guide. To this end, a corresponding guide bushing (not functioning as a seal) is used.

Metering Pumps. The chemical industry as well as the food processing industry needs metering pumps on a large scale to ensure fluid is metered precisely and consistently. Delivery pressures of up to 4000 bar and precisely repeatable metered quantities of 0.001 l/h or 0.2 mg/stroke up to 200 000 l/h have to be achieved. The required metered quantities are regulated by adjusting the speed or more frequently the plunger stroke. While continuously variable transmissions or frequency converters are widely used to adjust the speed, different variants are used to adjust the piston stroke, which, depending on their design, operate with the *fixed mean position of the piston* or with the *fixed dead center of the piston* from stroke zero onward.

Both plunger- and diaphragm-type metering pumps are constructed.

The stroke can be adjusted in different ways, e.g.:

- Crankshaft drive with adjustable crank radius (Fig. 10.50).
- Double eccentric with adjustable eccentricity (Fig. 10.51).
- Radially adjustable rotary eccentric (Fig. 10.52).
- Spring cam drive units (Fig. 10.53) are often used for simple, primarily smaller diaphragm metering pumps. An adjustable limiting bolt mechanically adjusts the stroke. The stroke can be adjusted by using the limiting bolt to increase or decrease the

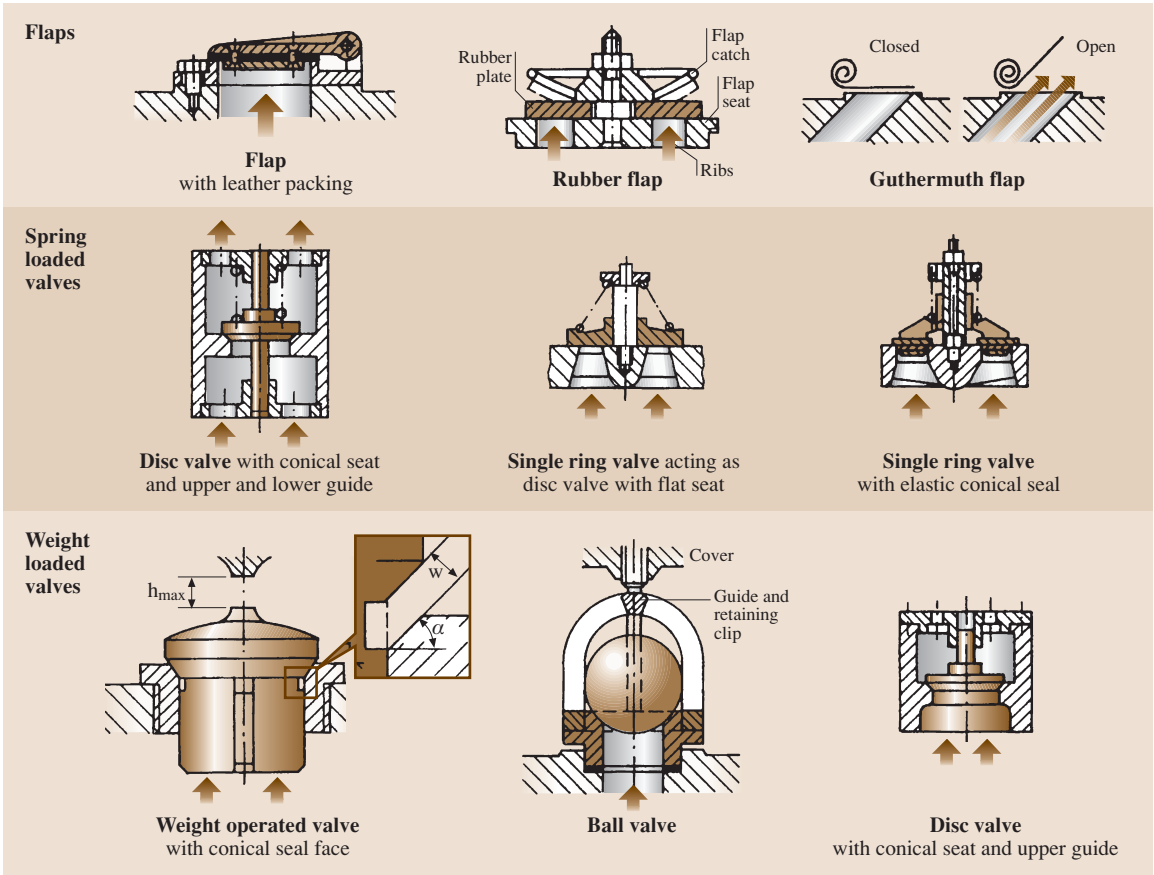


Fig. 10.46 Valve types (after [10.12])

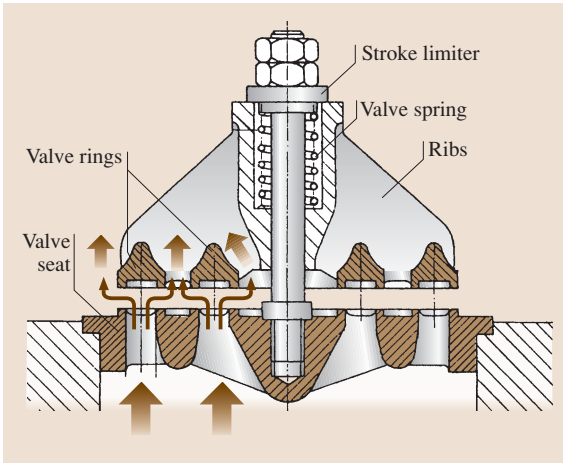


Fig. 10.47 Multiple annular valves (after [10.12])

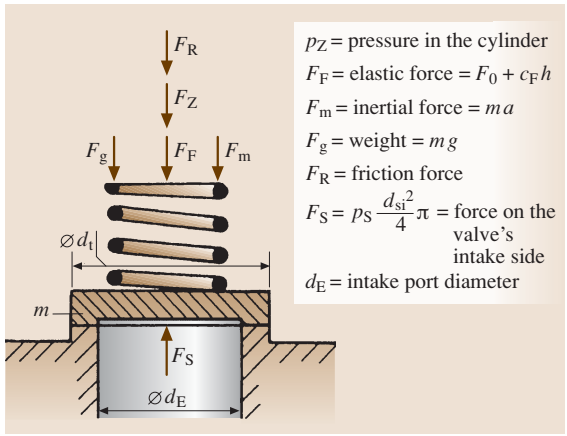


Fig. 10.48 Forces at the suction valve

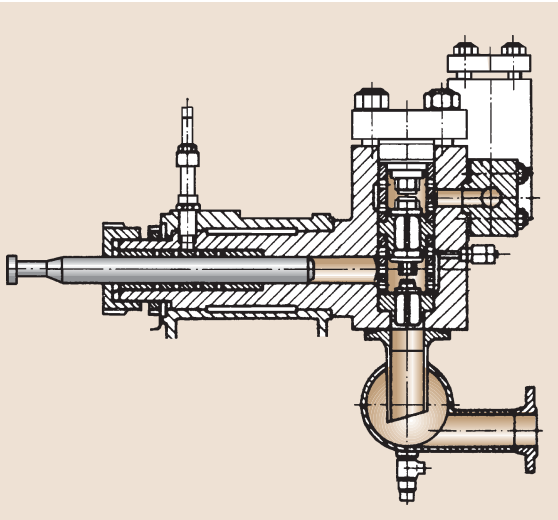


Fig. 10.49 Maximum pressure piston pump for pressures up to 7000 bar (after [10.13])

distance between the follower and the cam's basic circular path. A rotary eccentric cam is the cam shape used. A so-called phase angle occurs in which only part of the stroke is utilized. However, strong forces of acceleration act in this drive unit at the start of the effective stroke so that the range of adjustment at higher pressures is limited to values between a complete stroke and a maximum of 50% of the stroke.

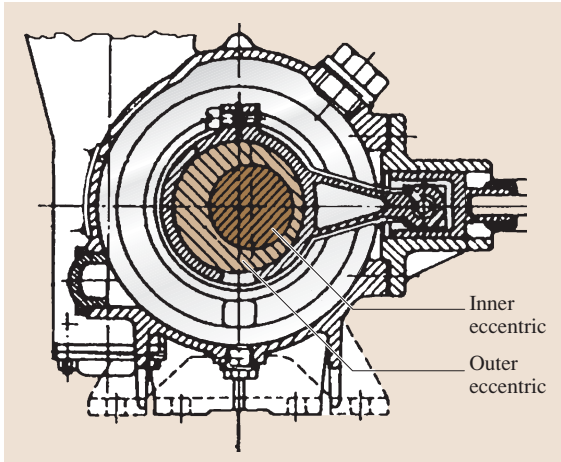


Fig. 10.51 Double eccentric (after [10.14])

Diaphragm metering pumps are used for applications in which the delivery chamber must be absolutely impermeable to the environment. The diaphragm can be operated mechanically as well as hydraulically (Figs. 10.54 and 10.55).

The delivery pressure in mechanical drives subjects the membrane edge to great stress. A hydraulic drive does not subject the diaphragm to great stress and considerably higher delivery pressures are possible.

With regard to adjusting the stroke, the same applies to hydraulically powered diaphragms as to mechanically powered diaphragms. The hydraulic system is

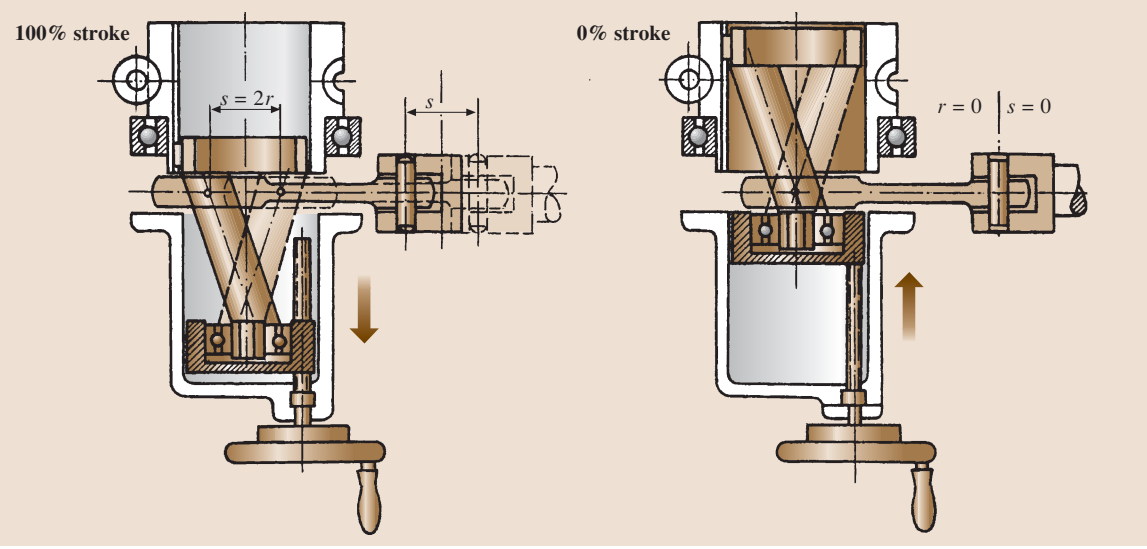


Fig. 10.50 Adjustable crank radius (after [10.11])

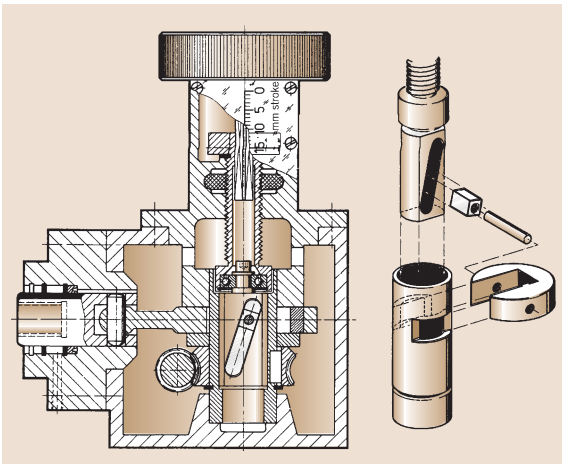


Fig. 10.52 Adjustable rotary eccentric (after [10.14])

simply connected between the plunger and the diaphragm and serves to transfer the quantity pumped in a ratio to the volume of 1 : 1. The force and travel ratios follow the continuity equation. The amount of leakage potentially occurring in the hydraulic system is replenished from the storage tank during each stroke so that the diaphragm's stroke motion is retained with its end positions. Frequently, there is a diaphragm position control. Employing a device to redefine the diaphragm's end position during each stroke prevents overloading of the diaphragm.

A multiple membrane with an inserted control medium is used for cases in which contamination of the pumped fluid by the hydraulic fluid must be prevented or emergency operation guaranteed.

Rotary Piston Pumps

Along with the type of motion, the absence of working valves is a significant feature of piston pumps with rotary displacers. Ports control the suction and pressure

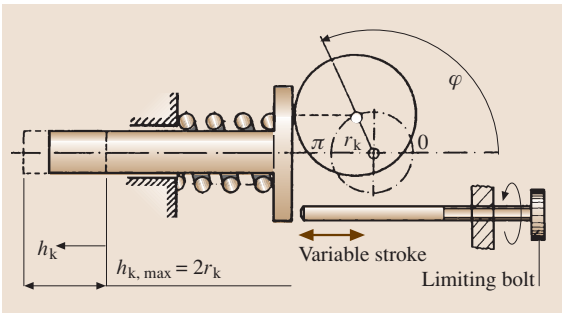


Fig. 10.53 Flexible cam drive unit

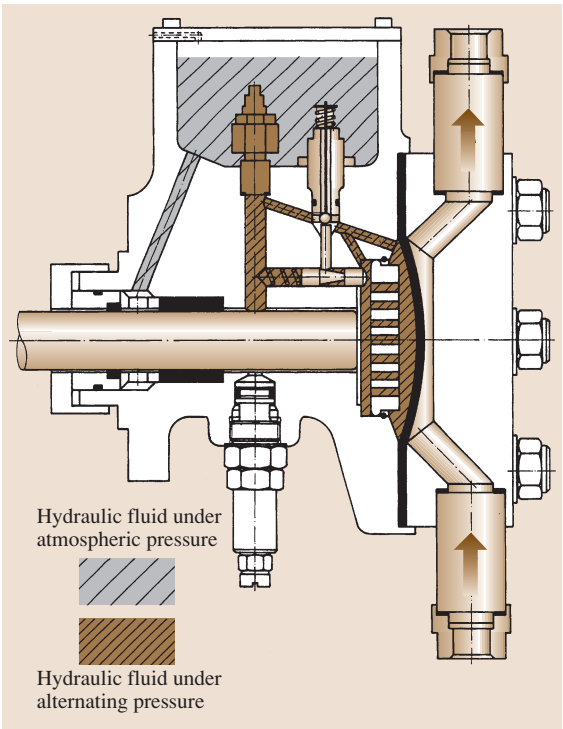


Fig. 10.54 Diaphragm metering pump with diaphragm position control (after [10.11])

process. As a result, when a pressure difference exists in the adjacent medium, these machines can also function as power machines. In order to prevent this, non-return valves (that do not have the function of working valves) are used.

Single Rotation Piston Machines.

Gear Pumps. A gear pump consists of two or more gearwheels arranged in a housing and intermeshed. The driveshaft is attached outside the housing, the remaining gearwheels being driven by the gearing. The suction and pressure sides are separated by the meshing and the sealing gap between the tooth tips, gear-side surfaces and housing.

The rotating gearwheels cause the pumped medium to be delivered from the suction side to the pressure side. In the zone of gear meshing, the teeth displace the medium from the tooth gaps (displacement effect). Simultaneously, the intermeshed tooth flanks seal the suction chamber off from the pressure chamber (Fig. 10.56 (Sp)). Moreover, the tooth tips seal the individual volumes off from the housing (Fig. 10.56). In addition, the axial gap must be confined to narrow lim-

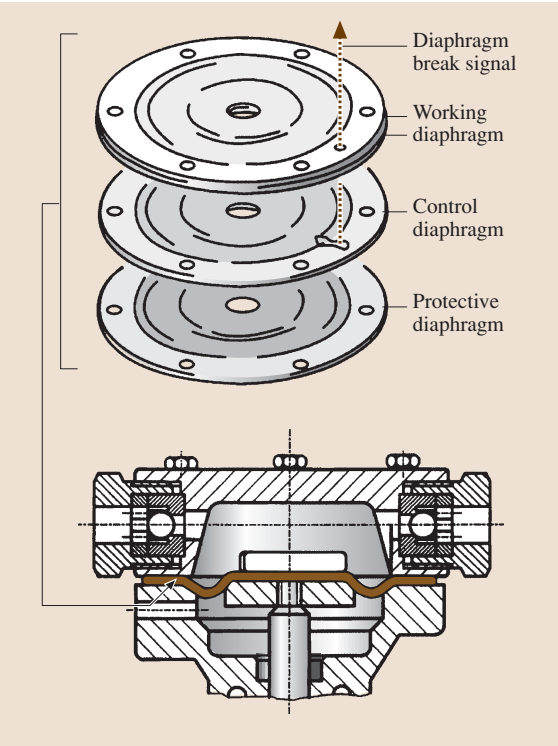


Fig. 10.55 Multi-diaphragm pump (after [10.11])

its by maintaining slight axial clearances. An optimal seal enables the gear pump to attain high delivery rates (volumetric efficiencies).

As a rule, the same involute profile used for transmission gears is used for the tooth flank profile of gear pumps. Thus, the gearwheels can be manufactured cost effectively by using the same tools on identical machines. Other, more-expensive profiles are reverted to only for special applications such as pulsation-free sine pumps. Helical gears are also rarely used. While intended to achieve smoother running, they generate an axial thrust to the gearwheels and bearings as well as unequal axial gap widths.

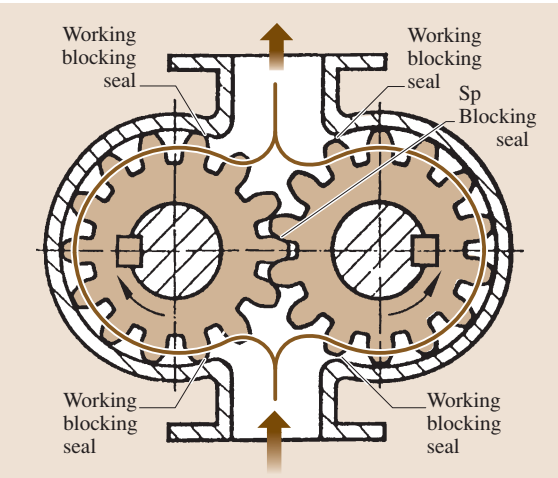


Fig. 10.56 Principle of the gear pump

When rolling, a self-contained volume is produced in the tooth gap into which the opposing tooth immerses. This causes a change of volume. Since a change of volume in enclosed spaces cannot be tolerated when fluid is being pumped, appropriate measures must ensure fluid is supplied to and removed from the tooth gap (Fig. 10.57).

To this end, either so-called compression slots are made in the side panels of the pump or on the teeth flanks or boreholes are made in the bottom land of a gear, which the grooves in a fixed axis connect to the suction or pressure chamber. Alternatively, the tooth flank clearances are made sufficiently large.

When they are being engineered, the clearances between the components must be dimensioned to correspond to the viscosity of the pumped medium and the delivery pressure. This can lead to problems, most notably when there is axial play.

Axial hydraulic clearance compensation is applied to improve impermeability at high pressures and to simultaneously reduce friction during starting. Thus, delivery pressure is utilized to reduce running clearance.

Internal gear pumps can additionally be given radial hydraulic clearance compensation (Fig. 10.58).

Low-pressure pumps ($\Delta p < 16$ bar) are usually designed with eight to 14 teeth while medium- to high-pressure pumps ($\Delta p > 160$ bar) have 16–25 teeth.

Single Rotation Piston Pumps. Rotational motion piston pumps are often incorrectly called eccentric rotational motion piston pumps. Since they are frequently positively driven by external gears, the sliding of their plungers past one another without contact is common to all rotational motion piston pumps. As a result, these pumps are well suited for delivering non-lubricating and abrasive liquids. This aspect fundamentally distinguishes this type of pump from gear pumps. The types are classified according to the number of impellers and the type of hub design (Fig. 10.59).

In some cases, the design selected for the displacer ranges from a multi-impeller to a gear type. However, the rotational motion piston pump is not a gear pump

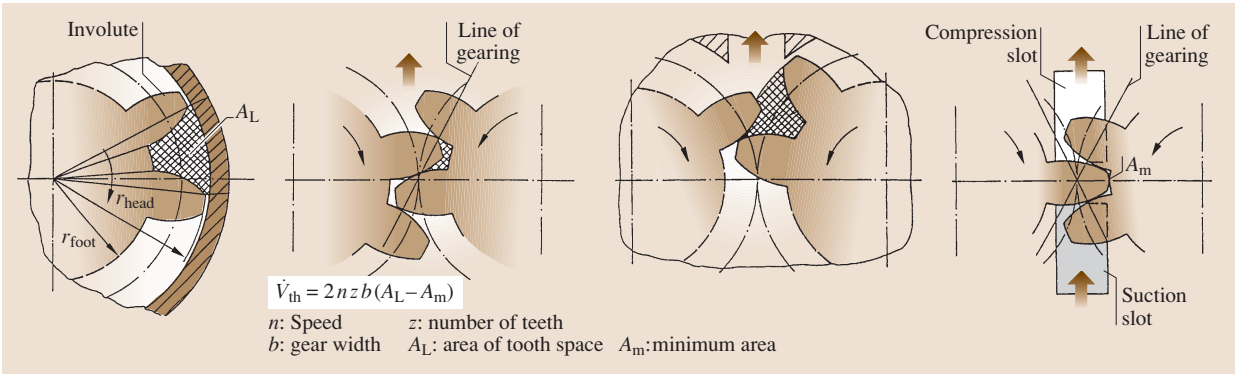


Fig. 10.57 Displacement in a gear pump

since exterior synchronizer gears enable the displacer to run without contact.

The fixed hub produces a seal covering the space between the displacers. The plunger shape creates a flat seal in the space between the housing and the displacer so that the fixed hub is able to achieve better inner impermeability.

These pumps are used, for instance, to deliver finishes, paints, oils, gasoline, food and alcohol and for foaming liquid–gas mixtures.

The pressure range of these pumps reaches 20 bar at delivery rates up to 500 m^3/h. They can pump media with very high viscosities.

Other displacer forms, some without dead space, are also used for special applications such as those in the food processing industry (Fig. 10.60).

Rotary Screw Pumps. Among other things, rotary screw pumps are called *degenerate* gear pumps with helical gears and few teeth. The inter-rotating profiles effect an axial delivery. Rotary screw pumps are constructed with two, three or five spindles. The design with two or three spindles is most widespread. The secondary spindles are arranged around the main spindle (Fig. 10.61).

These pumps are engineered both with single and dual flow to compensate for the axial thrust. The main spindle is usually double-threaded; the secondary spindles double or triple threaded. Depending on the pumped medium, pitch and profile, they are constructed with or without synchronizer gears.

Basically three profiles are used:

- Involute profile, without synchronizer gears (e.g., Leistritz)
- Epicycloidal profile (e.g., IMO AB, Stockholm, Sweden: manufacturer of rotary positive 3 screw pumps)
- Trapezoidal or rectangular profile, only with synchronizer gears (e.g., Bornemann)

Rotary screw pumps are constructed for:

- Delivery rates of 10–150 m^3/h
- Pressures of 15–150 bar
- Speeds of 500–1500 min^{-1}
- Viscosities of 25×10^{-6}–5×10^{-3} m^2/s

Planetary Rotation Piston Machine.

Eccentric Screw Pumps. The principle of the eccentric screw pump (Fig. 10.62) goes back to the Frenchman Moineau. A rotor with a single-threaded diamond knurl rotates in a housing (stator) with a double-threaded diamond knurl. An alternative name for the eccentric screw

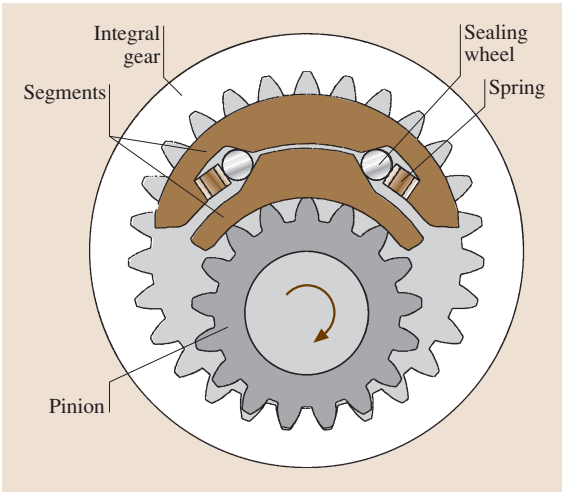


Fig. 10.58 Internal gear pump with clearance compensation (after [10.15])

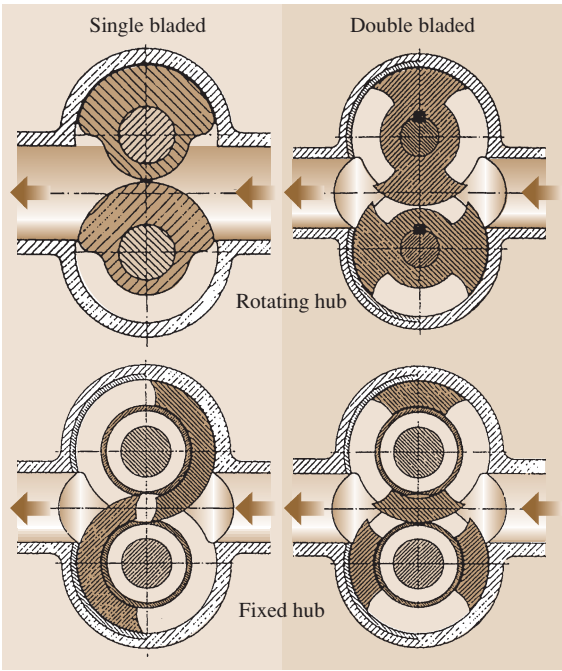


Fig. 10.59 Variants of rotational motion piston pumps (after [10.16])

pump is the single-spindle pump. Depending on its design, this pump is an eccentric rotational motion piston pump.

At every point, the rotor has a circular section with its center lying on the rotor axis on a helix. The housing has an oval section and twice the pitch of the rotor.

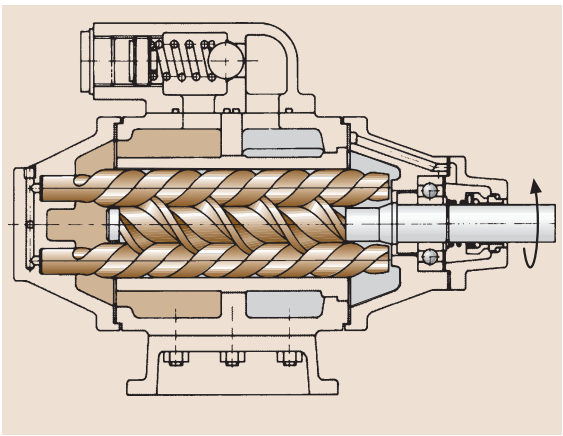


Fig. 10.61 Rotary screw pumps with three spindles (after [10.17])

Eccentric screw pumps attain delivery rates of 0.6–1000 m^3/h at pressures of up to 40 bar. Apart from being able to deliver abrasive and highly viscose media, the eccentric screw pump provides the advantage of a pulsation-free delivery rate (Fig. 10.63). Since dry running the pump has to be avoided in any operational situation, the suction and compression flanges in this type of pump are frequently arranged on the top side of the housing.

Typically, the rotors are made of CrNi steel or cast iron and the housing of red brass or cast iron. The stator is made of elastic material such as rubber, polytetrafluoroethylene (PTFE) or neoprene.

The medium may contain a solid content of up to 45%. However, this leads to a sharp power increase and

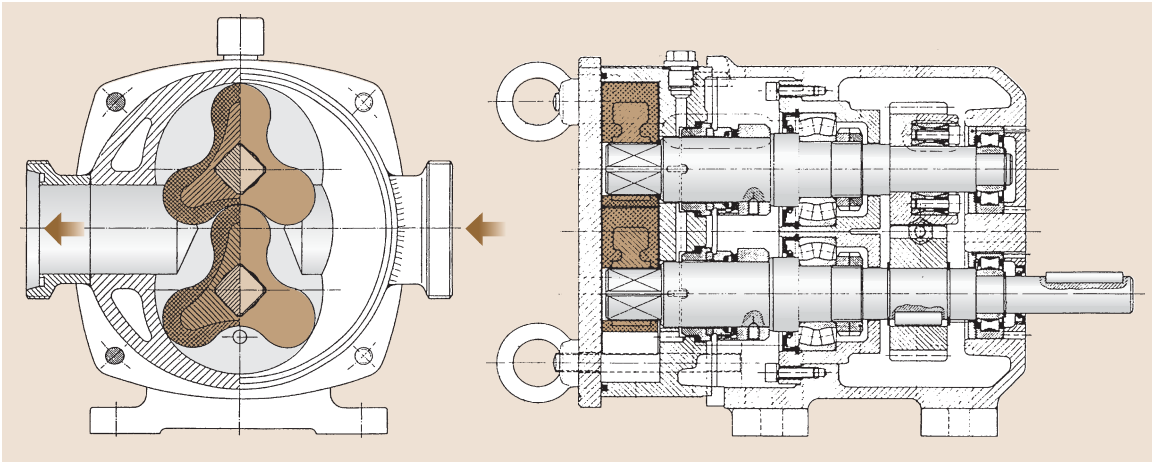


Fig. 10.60 Rotational motion piston pump for food, cross and longitudinal sections (after [10.16])

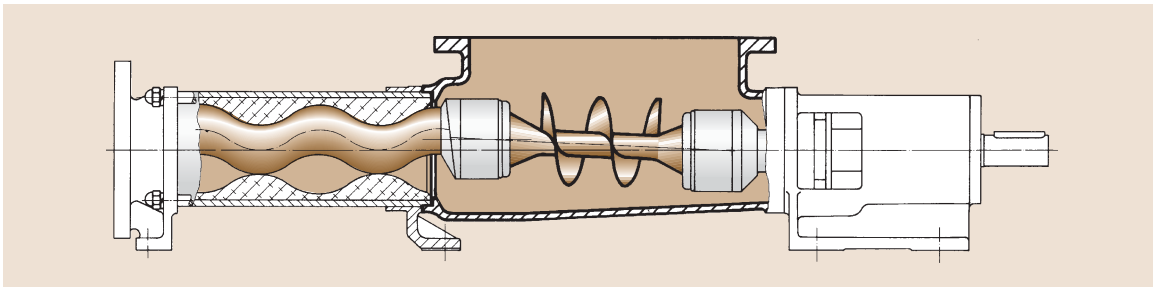


Fig. 10.62 Eccentric screw pump (after [10.18])

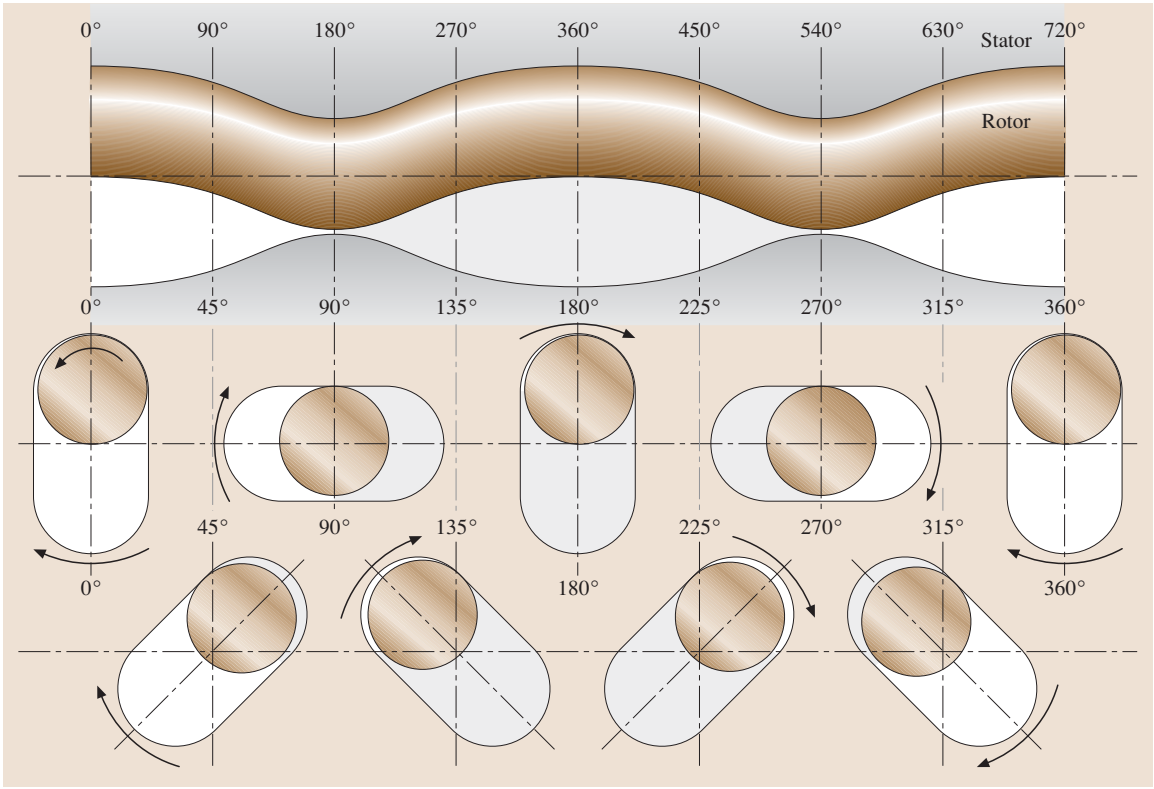


Fig. 10.63 Displacement effect of the eccentric screw pump (after [10.18])

causes low speeds. The elastic stator makes it possible to attain high impermeability and a suction head of up to 8.5 m.

The rotor executes a circular motion with its center axle so that it must be driven by cardan joints. In some cases, only one cardan joint is used in smaller units. As a result, the stator must absorb part of the motion. In addition, a joint can be used, which compensates for a parallel axle offset. Impellers or screws that improve the feeding of the pumped substance are often mounted on the cardan shafts.

Eccentric screw pumps are used to pump mortar, among other things.

Circulation Piston Machine.

Cell Pumps. Cell pumps (Fig. 10.64) are frequently used for low delivery rates of up to 25 m^3/h. As a rule, four to six slide valves are employed. Up to 12 slide valves are used at higher pressures of up to 100 bar. In some cases,

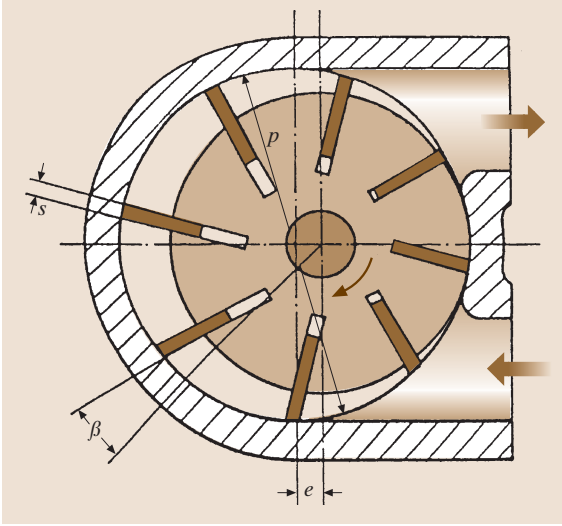


Fig. 10.64 Cell pump

even two slide valves per channel are used to increase impermeability.

Normally, the centrifugal force alone lifts the slide valves negatively. As a result, these pumps are relatively leaky at low speeds. The principle is broadly applied to fuel feed pumps for instance. The roller cell pump is another common type. Rollers are used instead of slide valves. Their shape gives them advantages over slide valves in terms of wear.

This type of pump is used for example as an electric fuel pump to pump gasoline in cars (Fig. 10.65).

So-called impeller pumps are also used when delivery flows are smaller, e.g., for garden pumps or coolant pumps for boat engines (Fig. 10.66).

The working chamber changes volume by bending the rotor's elastic, usually rubber impellers against the sickle-shaped part of the housing deviating from a circular shape.

The rotor is installed axially without any gap in order to guarantee high impermeability. Thus, these

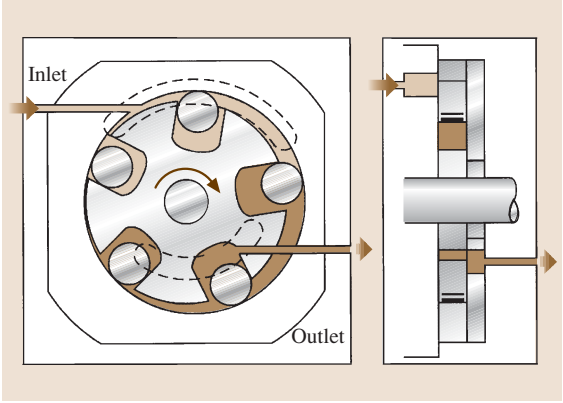


Fig. 10.65 Roller cell fuel pump (after [10.19])

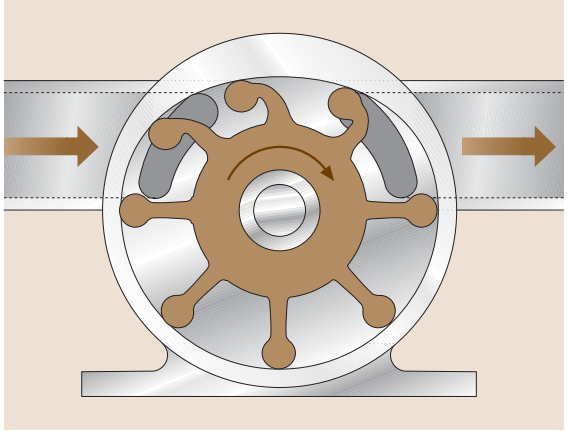


Fig. 10.66 Impeller pumps (after [10.20])

pumps are self-priming. The impellers' ability to upright themselves determines the maximum speed. The delivery pressure depends on the flexural strength of the impellers. These pumps are insensitive to contamination of the pumped medium but can only reach low pressures of approximately 4 bar. They are used as boat motor seawater pumps and as garden pumps.

10.3 Compressors

Compressors can be divided into two general categories: positive displacement type and non-positive displacement types. Positive displacement compressors deliver essentially the same volume of air regardless of the pressure ratio while non-positive displacement compressors will have reduced flow at higher delivery pressures. Axial flow and centrifugal compressors are non-positive displacement types and reciprocating, rotary-screw, and diaphragm compressors are of the positive displacement type. This section discusses the characteristics of positive displacement reciprocating compressors.

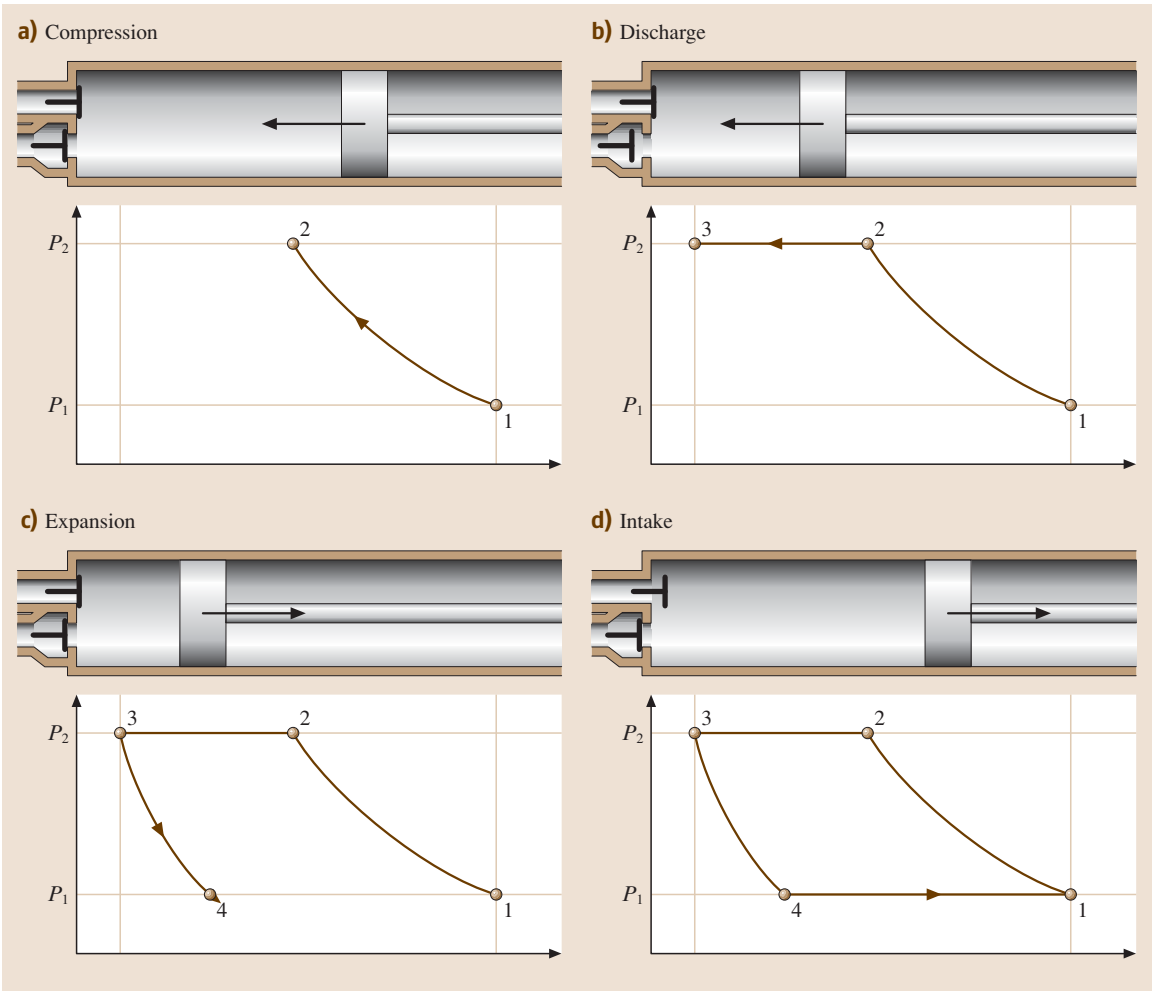


Fig. 10.67a–d Reciprocating compression processes: **(a)** compression **(b)** discharge **(c)** expansion **(d)** intake

Reciprocating compressors are frequently used when high pressures are needed, although they generally use multiple stages of compression to limit the power required for the process. The basic compression process can be visualized as occurring in a piston–cylinder configuration that is similar to that of an internal combustion engine. In practice, reciprocating compressors frequently use both sides of the piston in *double-acting* systems. Inlet and discharge valves are spring-loaded so that they will automatically open when there is an appropriate pressure in the cylinder. The inlet valve will open when the cylinder pressure drops below the inlet pressure, and the discharge valve will open when the cylinder pressure exceeds the receiver pressure.

10.3.1 Cycle Description

Figure 10.67 shows the sequence of processes followed by reciprocating compressors. The upper part of each diagram shows the piston position in the cylinder and whether the valves are open or closed. The lower part of each diagram shows the corresponding pressure–volume diagram for the process. In Fig. 10.67a the piston moves to decrease the volume with the valves closed. The pressure rises from state 1 to state 2 following a process line that depends on the amount of heat loss during the compression process. When the cylinder pressure exceeds the receiver pressure, the discharge valve will be forced open and the gas is discharged at constant pressure, as shown in Fig. 10.67b. When the

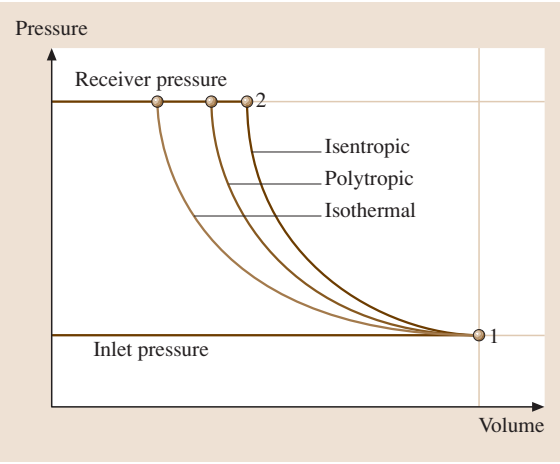


Fig. 10.68 Compression process lines

piston reaches the point of minimum volume, the cylinder volume is equal to the clearance volume. This is the starting point for the expansion process from state 3 to state 4, shown in Fig. 10.67c, where the pressure falls from the receiver pressure to the intake pressure. Finally, when the cylinder pressure has dropped to the inlet pressure, the intake valve opens and fresh air is drawn into the cylinder as shown in Fig. 10.67d.

The compression process for a reciprocating compressor will generally follow a polytropic process line that is between a reversible adiabatic process (isentropic) and a constant temperature (isothermal) process as shown in Fig. 10.68. The isothermal process accomplishes the compression with the least work input. Ideal gases follow the relationship

$$pV = \mathrm{R}T \,. \tag{10.62}$$

When an ideal gas is compressed it usually follows a polytropic process line that is modeled as

$$pV^n = \text{const} \,. \tag{10.63}$$

The usual values for n are between 1, which corresponds to an isothermal process and k, which is the ratio of specific heats and is equal to 1.4 for air at 25 °C.

For the polytropic process, the pressure at the end of the compression process will depend on the compression ratio $r_\mathrm{v} = \frac{V_1}{V_2}$ according to the following equation

$$p_2 = p_1 \left(\frac{V_1}{V_2}\right)^n = p_1 r_\mathrm{v}^n \,. \tag{10.64}$$

The final temperature can be calculated as

$$T_2 = T_1 r_\mathrm{v}^{n-1} = T_1 \left(\frac{p_2}{p_1}\right)^{\frac{n-1}{n}} = T_1 r_\mathrm{p}^{\frac{n-1}{n}} \,. \tag{10.65}$$

The power required for the compression process is

$$Pwr = p_1 \dot{v}_1 \left(\frac{n-1}{n}\right)\left(r_\mathrm{p}^{\frac{n-1}{n}} - 1\right) \,, \tag{10.66}$$

where $\dot{v}_1$ is the volumetric flow rate entering the compressor, r_p is the pressure ratio p_2/p_1, n is the polytropic exponent, and p_1 and p_2 are the pressures at states 1 and 2, respectively.

At high pressures, many gases will deviate considerably from ideal gas behavior. Usually, these deviations are incorporated into the calculation through the use of a compressibility factor Z. The equation between temperature, pressure and volume becomes

$$pV = ZRT \,. \tag{10.67}$$

The value of Z depends on the state, and charts are widely available for most common gases.

While reciprocating compressors are usually designed to encourage heat rejection, it is more common to have an intercooler between separate stages of compression. The intercoolers and multiple compressor stages are usually integrated into a common assembly that may take advantage of double-acting pistons.

10.3.2 Multi-Staging

Compression at constant temperature requires less work input than when the temperature is allowed to rise. In practice, isothermal compression is difficult to achieve. However, incorporating intercooling between multiple stages of compression can approximate isothermal compression.

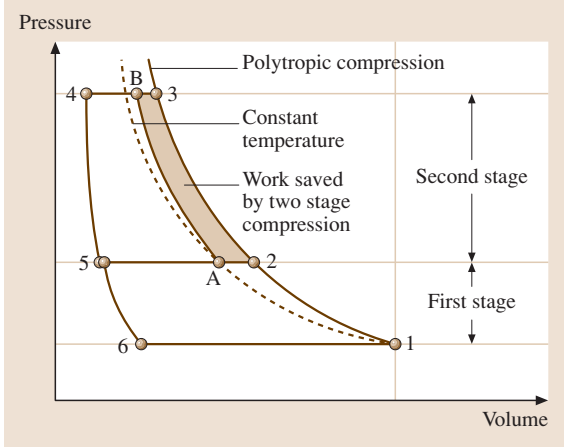


Fig. 10.69 Two stage compression

In the limiting case of an infinite number of compression stages that include intercooling back to the inlet temperature, the total work required will be equal to that for isothermal compression. Even the division of a compression process into two stages can save considerable work input.

A two-stage compression process is depicted in Fig. 10.69. The compression process begins at state 1, which corresponds to the piston at BDC and inlet pressure. The gas is compressed in the first stage to the inter-cooler discharge pressure at state 2. The gas is then discharged at p_2 into an intercooler before entering the second compressor stage at state A. In this idealized case, the intercooler is assumed to operate at constant pressure but the volume of gas is decreased due to the drop in temperature to the initial temperature of the gas. Then, the gas is compressed to state B before being discharged at p_B. If the pressure increase had been attempted with a single stage of compression, then the process line would pass through states 1, 2, and 3. Since the area enclosed on the p–V diagram is the work needed to accomplish the process, the shaded area is the difference in work between the single stage and two-stage compression processes. Clearly, the two stage compression is more efficient.

The optimum pressure for intercooling is generally assumed to correspond to an equal pressure ratio for each stage. This assumes the intercooling is able to reduce the gas temperature to the inlet temperature at each stage. If r_s is the pressure ratio for each stage, r_t is the overall pressure ratio, and s is the number of stages, then

$$r_s = \sqrt[s]{r_t} \,. \tag{10.68}$$

10.3.3 Design Factors

At the end of the compressor's discharge stroke, shown in Fig. 10.67b, gas fills the clearance volume at the discharge pressure. This gas expands as the piston moves away from the cylinder head until its pressure drops below the inlet pressure when the intake valve opens. The induction of fresh charge does not begin until this point is reached so the full volume displaced by the piston is not utilized. When the clearance volume is large, then the capacity of the compressor is less.

The volumetric efficiency of a compressor can be approximated as

$$\eta_v = 1 - \frac{V_{\text{clearance}}}{V_{\text{Displ}}}\left(r_{\text{p}}^{\frac{1}{k}} - 1\right) - \text{Leakage} \,, \tag{10.69}$$

where the leakage can generally be assumed to be between 0.03 and 0.05. Lower-molecular-weight gases usually have higher leakage.

It can be seen from the equation for η_{v} that an increase in clearance volume directly causes a decrease in volumetric efficiency. Its significance is much greater for high values of the pressure ratio r_{p}.

Although compressor designers try to minimize it, the clearance volume cannot be entirely eliminated. Reducing it to less than 4% of the displacement volume is difficult. The amount of clearance volume will affect the capacity of the compressor and its efficiency. Overall compression efficiency is improved when valve flow area is large. However, the desire to minimize the clearance volume conflicts with the desire to maintain large valves. Thus, there is often a tradeoff between volumetric efficiency and compression efficiency, which determines the actual value of the clearance volume.

10.4 Internal Combustion Engines

In an internal combustion engine, the working fluid consists of a fuel–air mixture and the combustion products of this mixture. Although many cycles have been proposed, the traditional two-stroke and four-stroke cycles still dominate current use.

Depending on their design and application, internal combustion engines provide excellent portability, power density, and fuel economy. Vehicles that utilize internal combustion engines provide unsurpassed range, drivability, and driver comfort while maintaining low levels of hazardous pollutants.

10.4.1 Basic Engine Types

Engines can be categorized in many different ways. The number of cylinders, the type of valve actuation, and whether the engine is turbocharged or naturally aspirated are all possible choices. Some engines are spark-ignited and utilize a homogeneous fuel–air mixture and some are compression-ignited, also called diesel engines, and utilize a heterogeneous fuel–air mixture. Another characteristic is whether the engine uses the two-stroke or four-stroke engine cycle.

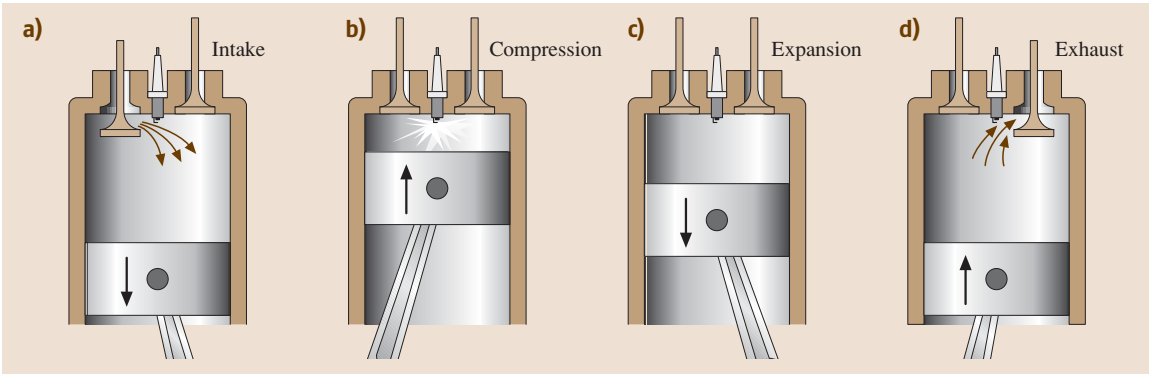


Fig. 10.70a–d Four stroke cycle events

All reciprocating internal combustion engines need to go through the four processes of intake, compression, expansion, and exhaust. The basic difference between two-stroke and four-stroke cycle engines is that the two-stroke engine accomplishes the four processes in a single revolution, or two strokes of the piston (one up and one down). The four-stroke engine needs two revolutions to complete the cycle.

The processes are shown in Fig. 10.70 for a four stroke cycle. In Fig. 10.70a, the intake valve is open and fresh charge is drawn in as the piston moves downward toward the bottom dead center (BDC) position. After BDC, the intake valve closes and the piston compresses the air on its upward compression stroke. Near the end of the compression stroke, close to top dead center (TDC), the spark plug fires and ignites the fuel–air mixture. In a diesel engine, only air is drawn in through the intake valve and fuel is injected into the air near the end of the compression stroke. This fuel self-ignites after a short delay. For both spark-ignited and diesel engines, the combustion products do work on the piston as it moves out for the expansion stroke as shown in Fig. 10.70c. In Fig. 10.70d, the exhaust valve opens near BDC and the combustion products are expelled from the cylinder by the upward motion of the piston. At the end of the exhaust stroke, the intake valve opens and the cycle is repeated with another intake stroke.

The two-stroke cycle is depicted in Fig. 10.71. When the piston is near the BDC position, air enters the cylinder from a port in the cylinder wall. There is a deflector on the top of the piston to inhibit the direct passage of the fresh charge across the cylinder and out the exhaust port on the other side. As the piston moves upward, it covers the intake and exhaust ports and compresses the fuel–air mixture. Near TDC, the spark fires to start the combustion process. In a two-stroke diesel engine the fuel is injected into the compressed air at this point. For both types of engines, the combustion products expand and do work on the piston surface until the point where the piston uncovers the upper edge of the exhaust port. At this point, the gases in the cylinder rapidly blow down until the pressure in the cylinder

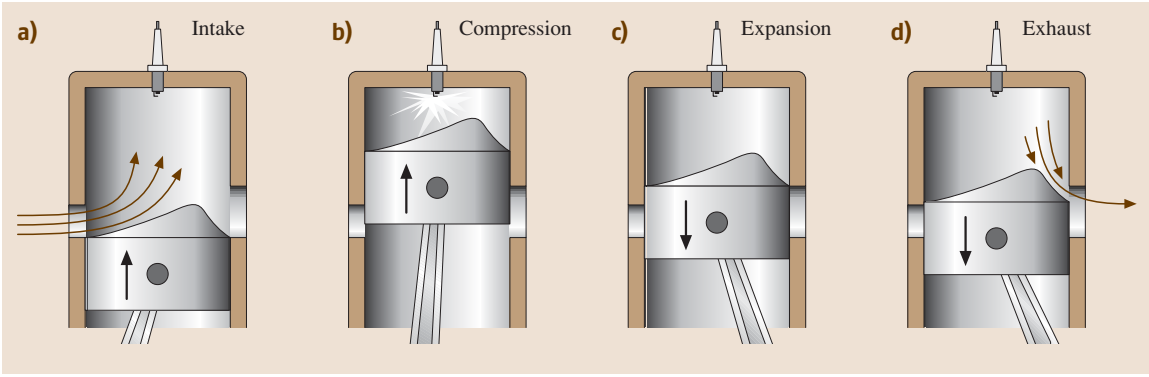


Fig. 10.71a–d Two stroke cycle events

is close to the exhaust pressure. As the piston continues its downward motion, it uncovers the intake port so fresh charge can enter the cylinder and repeat the cycle.

Because the two-stroke engine needs to complete all four processes in a single revolution, it must start the exhaust process well before the piston has reached bottom dead center. A four-stroke engine can wait until about 140° after TDC before starting to open the exhaust valve since the primary criterion is that the period of rapid pressure equalization that occurs when the valve is first opened, called the blowdown, is essentially complete by BDC. A two-stroke engine must start opening the exhaust port at about 90° after TDC to provide sufficient time for the blowdown before the intake port opens and the intake process starts. Opening the exhaust valve early so that the cylinder pressure is throttled down to ambient allows no recovery of the energy in those hot gases and is the major reason why two-stroke engines are less efficient than four-stroke engines.

In four-stroke engines, almost the entire cylinder contents of burned product gases, called residual gases, is expelled when the piston reaches TDC at the end of the exhaust stroke. Then, as the piston moves away from TDC, it produces a low pressure in the cylinder which draws in a fresh air charge through the intake valve. Four stroke engines are said to be self-scavenging. That is, the piston motion is directly responsible for moving the exhaust gases out of the engine and drawing in fresh air. Two stroke engines are not self-scavenging. Some mechanism other than piston motion is needed to exchange the gases in the cylinder. By opening the exhaust valve early, while the cylinder pressure is still high, most of the residual gases can be expelled. The fresh charge must be forced into the cylinder from a pressurized source, which might be a crankcase that is pressurized by the downward motion of the piston as in single cylinder two-stroke engines used for hand-held power equipment. It might also be from an engine-driven blower as is common in diesel two-stroke engines.

10.4.2 Performance Parameters

Engine speed and torque are the two most fundamental quantities of engine performance. Both are usually measured quantities with speed given in revolutions per minute (rpm) and torque in Newton-meters (N m). Power is defined as torque times rotational speed. When the torque is measured at the engine flywheel, it is called the *brake torque* and the power calculated from it is the brake power, as

$$P_b = T_b \omega = T_b 2\pi N \,, \tag{10.70}$$

where P_b is the brake power, T_b is the brake torque, ω is the rotational speed, usually in radians/unit time, and N is the rotational speed in rev/unit time.

The source of power in the engine is the work done on the piston by the expanding combustion gases. The power associated with this piston work is called the *indicated power*. The difference between the indicated power and the brake power is the power required to overcome friction and to drive the accessories including the water and oil pumps

$$P_i = P_b + P_f \,, \tag{10.71}$$

where P_i is the indicated power, P_b is the brake power, and P_f is the friction power.

Direct calculation of the indicated power requires measurement of the cylinder pressure. The brake power can be calculated from the measured engine torque and speed. The friction power must be calculated from the difference of the indicated and brake power.

Generally, small engines run at high rotating speeds and large engines run at low rotating speeds. To allow comparisons between engines of different sizes, it is common to calculate the *mean piston speed*, which is the average velocity of the piston as the engine makes one revolution

$$\text{MPS} = \frac{2S}{\text{time for one revolution}} = 2SN \,, \tag{10.72}$$

where S is the stroke, and N is the engine rotating speed.

The *mean effective pressure* (MEP) is a way to normalize the work done by the engine against the size of the engine. It is intended as a measure of engine loading. The MEP is defined as the ratio of the work done by the engine in one cycle to the displacement volume. A four-stroke engine undergoes one cycle in two revolutions (or 4π radians) so the work done is equal to the torque times the angular displacement. When the brake torque is used, the quantity is known as the *break mean effective pressure* (BMEP)

$$\text{BMEP} = \frac{4\pi T_b}{V_d} \,. \tag{10.73}$$

Since work is measured in N m and volume in m^3, the ratio has units of pressure N/m^2. The MEP is sometimes described as the pressure which, if applied as a constant pressure during the expansion stroke, would give the same work as actually produced by the engine.

The *mechanical efficiency* is a measure of how much of the power produced by the combustion process is delivered to the output shaft. It is defined as the ratio of the brake power to the indicated power

$$\eta_m = \frac{P_b}{P_i} = 1 - \frac{P_f}{P_i}, \tag{10.74}$$

where η_m is the mechanical efficiency, P_b is the brake power, P_i is the indicated power, and P_f is the friction power.

The *thermal efficiency* is defined as the ratio of the power produced by the engine to the rate at which fuel energy is supplied to the engine, as indicated by the lower heating value (LHV). When the brake power is used, the quantity is known as the brake thermal efficiency

$$\text{Brake thermal efficiency} = \eta_{bt} = \frac{P_b}{\dot{m}_{fuel}\,(\text{LHV})}. \tag{10.75}$$

The mechanical and thermal efficiencies are sometimes confused. For a modern engine running at full load, the mechanical efficiency may be 90% or higher. However, the thermal efficiency will generally be 30–45%.

The *specific fuel consumption* (SFC) is the ratio of the fuel flow rate to the power of the engine. When brake power is used, the quantity is known as the brake specific fuel consumption

$$\text{BSFC} = \frac{\dot{m}_{fuel}}{P_b}. \tag{10.76}$$

The BSFC is similar to an efficiency in that it measures how little fuel may be required to do a certain quantity of work. The lower the BSFC, the more efficient the engine.

The *volumetric efficiency* is a measure of how well air moves through the engine. For a four-stroke engine

$$\eta_v = \frac{\dot{m}_{actual}}{\dot{m}_{ideal}} = \frac{\dot{m}_{actual}}{\rho_{ref} V_d \left(\frac{\text{rpm}}{2}\right)}, \tag{10.77}$$

where η_v is the volumetric efficiency, $\dot{m}_{actual}$ is the actual mass flow rate of air (or air–fuel mixture) entering the engine, $\dot{m}_{ideal}$ is the ideal mass flow rate of air (or air–fuel mixture), ρ_{ref} is a reference density, V_d is the displacement volume, and $\frac{\text{rpm}}{2}$ is the number of engine cycles per minute.

For spark-ignited engines, the values of $\dot{m}_{actual}$ and $\dot{m}_{ideal}$ refer to the fuel–air mixture entering the engine. For diesel engines they refer only to the air entering the engine.

The volumetric efficiency tends to be ambiguous for several reasons:

1. There is uncertainty about where the reference density should be calculated. Some sources suggest using ambient conditions while others suggest using the average intake manifold pressure and temperature.
2. Although it is considered to be an efficiency, there is no reason why the volumetric efficiency cannot be greater than one. If the ambient density is used to compute the volumetric efficiency of a turbocharged engine, the volumetric efficiency may be as high as 2 or 3. Even a naturally aspirated engine with a tuned intake system can have a volumetric efficiency of 1.2 or 1.3.
3. In engines with a large valve overlap period, a significant amount of air can blow through the engine directly from the intake to the exhaust without participating in a combustion process. This air could contribute to a high volumetric efficiency but is not available for combustion.

10.4.3 Air Systems

The power produced by all internal combustion engines is limited by their ability to draw air from the atmosphere. Fuel systems can always be designed to provide the amount of fuel appropriate to this air flow. To increase the power produced by an engine of a certain displacement volume, the air flow needs to be increased. This can be done with passive techniques that are incorporated into the engine's design or through the addition of external devices, such as a supercharger.

Natural Aspiration

In a naturally aspirated engine, the air flows into the cylinder through the intake manifold and intake ports without the use of an external compressor of blower. At high engine speed this air moves at high velocity. When the piston approaches the end of the intake stroke, the momentum of the air keeps the air moving toward the cylinder and can continue to force air into the cylinder after the piston has started upward on the compression stroke. By properly timing the closing of the intake valve, the amount of air trapped in the cylinder can be increased beyond that which would be predicted based on ambient air density. This phenomenon, called the *ram effect*, increases with air velocity and therefore with engine speed.

A second effect that can be utilized to increase engine air flow is to take advantage of the pressure waves that are induced in the intake and exhaust system due to valve opening and closing events and piston motion. For example, the high-pressure wave created when the exhaust valve opens and rapidly blows down the cylinder contents travels to the end of the exhaust pipe and is reflected as a low-pressure wave or rarefaction wave. If this wave is timed to enter the cylinder near the end of the exhaust stroke it can assist in evacuating the residual gases and draw in fresh charge as the intake valve opens. Similarly, rarefaction waves in the intake system are reflected from the open end of the intake as pressure waves that will force more air into the cylinder. This process, known as tuning, is highly dependent on the relationship between valve timing, pipe lengths, and the speed of sound in the intake and exhaust gases. As a result, the benefits of tuning tend to be concentrated at specific engine speeds and the effects at other speeds may actually be negative. Other passive effects involving resonator cavities connected to the intake and exhaust pipes can also be used to raise the air flow to the engine. The engine air flow can also be increased with the addition of a compressor in a technique known as supercharging to be discussed later.

Effect of Speed on Volumetric Efficiency

Engine volumetric efficiency is affected by engine speed. At higher speeds, the pressure drop resulting from frictional effects associated with higher flow velocity tends to cause less air–fuel mixture to enter the cylinder. Up to a certain speed, this effect can be offset by the ram effect and tuned intake and exhaust pipes. These flow enhancing effects can keep the volumetric efficiency relatively high over most of the engine's operating range but at a certain speed, which depends mostly on valve area and mean piston speed, the volumetric efficiency drops off sharply.

Poppet Valves

Flow through a poppet valve is usually modeled as the product of isentropic flow through a restricted area passage and a flow coefficient that varies with valve lift and geometry. This equation is valid through all flow ranges, except when the flow is choked

$$\dot{m} = C_D \rho_0 c_0 A_x \times \sqrt{\left(\frac{2}{k-1}\right)\left[\left(\frac{P_x}{P_0}\right)^{\frac{2}{k}} - \left(\frac{P_x}{P_0}\right)^{\frac{k-1}{k}}\right]}, \tag{10.78}$$

where $\dot{m}$ is the actual mass flow rate through the valve, C_D is the discharge coefficient for the valve, ρ_0 is the density at the upstream stagnation state, c_0 is the speed of sound at the upstream stagnation state, A_x is the minimum flow area between the valve and seat, k is the ratio of specific heats $\frac{C_p}{C_v}$, P_x is the downstream static pressure, and P_0 is the upstream stagnation pressure,

C_D is greatest at low valve lifts. Under these conditions the air flow fills the entire flow channel without the separations that reduce the effective flow area. As the valve opens further, the flow separates from the valve edge and the seat, which reduces the actual flow area and decreases C_D. The value of C_D tends to be independent of the flow Reynolds number, except at low lift where C_D decreases as the Reynolds Number increases reflecting a greater significance of boundary layers on the flow.

Valve Timing

The timing for the opening and closing of the intake and exhaust valves can have a major impact on the engine performance. In four-stroke engines, the exhaust valve opens about 120–140° after TDC. Earlier timing will reduce expansion work and later may delay the exhaust blowdown so that high cylinder pressures early in the exhaust stroke increase the pumping work. Intake valve closing typically occurs at 20–60° after BDC and has a strong effect on the engine's volumetric efficiency. Earlier closing may reduce the air flow into the cylinder due to the ram effect described earlier, especially at high engine speeds. Later closing delays the start of the compression process and may allow some of the fresh charge to backflow into the intake manifold. The timings for closing the exhaust valve and opening the intake valve are not as critical for engine performance.

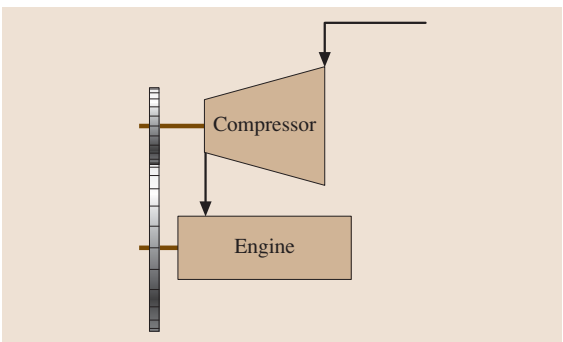


Fig. 10.72 Engine-driven supercharger

Supercharging

Supercharging is a general term for a variety of techniques used to boost the pressure of the air entering the cylinder to increase the engine's power. In some cases this involves devices to compress the air that might be driven directly by the engine or by the engine's exhaust gas. These devices are known as superchargers. When the device is driven by exhaust gas, it is called a turbo-supercharger or just turbocharger.

Figure 10.72 shows a schematic of an engine-driven supercharger that is driven by a set of gears, although belts are also frequently used. Engine-driven superchargers have the advantage that they provide air flow even at low speeds so they can be used to provide the scavenging needed for two-stroke engines during start-up. They also do not have the acceleration lag often noted with exhaust-driven turbochargers.

Turbocharging

A specific type of supercharging that utilizes power recovered from expanding the exhaust gas to atmospheric pressure is called *turbocharging*. Figure 10.73 shows a schematic of a typical system. Air, after being filtered, is compressed and then supplied to the engine. Single-stage radial-design compressors are most common and are generally capable of 3 : 1 pressure ratios. The compressor is directly coupled to an expansion turbine that provides the work to drive the compressor. While the turbine imposes a back pressure on the engine, this is more than offset by the higher potential power available from the increased pressure in the intake system.

In diesel engines, the air supplied by the turbocharger is more dense and allows more fuel to be injected while still maintaining the air–fuel ratio limits imposed by exhaust emissions concerns. In fact, due to the greater availability of air, turbocharged engines can usually be operated at higher air–fuel ratios than naturally aspirated engines which improves both particulate and NO_x emissions. Although engine power can be increased by a factor of 2–3 over a similarly sized naturally aspirated engine, fuel economy improves only slightly. This improvement is generally associated with the fact that friction losses do not increase proportionately with the increase in overall power so the mechanical efficiency of the engine improves. Turbocharging offers additional benefits of improved scavenging and piston cooling by allowing a portion of the compressed charge to short circuit through the cylinder during the valve overlap period.

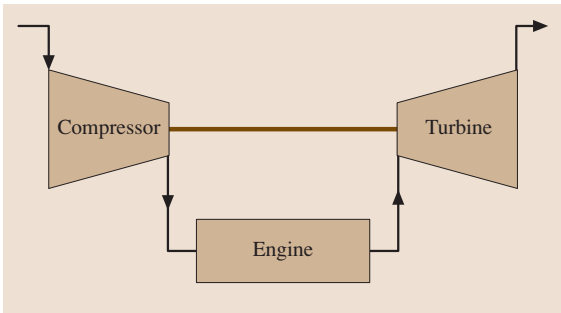


Fig. 10.73 Typical turbocharger configuration

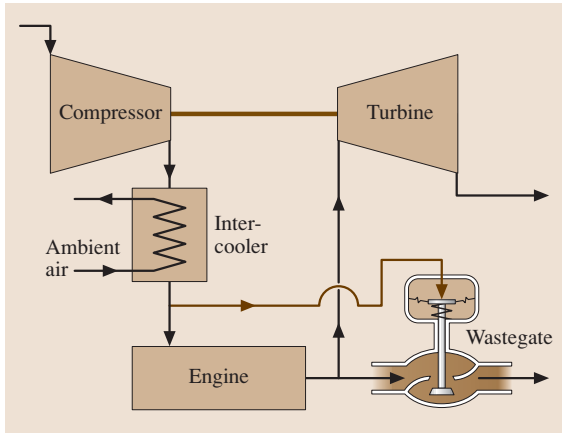


Fig. 10.74 Turbocharged engine with intercooler and waste gate

Part of the density increase potential of the turbocharger is lessened by the fact that the air leaving the compressor is at high temperature. It is common to insert a heat exchanger, called an intercooler or aftercooler, into the airstream after the compressor. Figure 10.74 shows this configuration. The compressed air may be cooled by engine coolant, or more commonly in heavy-duty applications, by ambient air.

Also shown in Fig. 10.74 is a waste gate. This device allows a portion of the engine's exhaust gas to bypass the turbine. A diaphragm actuator senses compressor boost pressure and releases a portion of the exhaust gas so that the boost pressure does not become excessive. Turbocharger design and matching to a specific engine is usually a compromise between providing sufficient air at the low-speed peak-torque condition while not delivering excessive pressure at the high-speed high-load conditions that will cause high peak cylinder pressures.

For engines that are consistently operated at high loads, there is more energy available in the exhaust gas than is needed to operate the compressor. One option

to utilize this excess energy is to add a second turbine, sometimes known as the power turbine, that will recover energy from the exhaust gas leaving the first turbine, which drives the compressor. As shown in Fig. 10.75, the power turbine is connected to the engine's drive train through a mechanical connection so the power can be delivered to the engine's output shaft. This technique is often referred to as *turbocompounding*. The fuel economy benefits of turbocompounding can be significant, especially when the engine is designed for minimum heat rejection, which tends to increase exhaust energy. However, at current fuel prices, the savings in operational cost has not been enough to justify widespread acceptance of this technology.

Other technologies have been developed to provide optimum turbocharger performance over a larger fraction of the engine's operating range. Fig. 10.76 shows a twin-turbocharger technology developed by Opel that utilizes two turbochargers. As shown in the figure, exhaust gas from the engine is directed through two separate passage-ways to two turbochargers. The passage to the larger turbocharger is equipped with a flapper valve to limit the flow. At light load conditions, the flapper is mostly closed, which forces most of the exhaust flow through the small turbocharger so it has sufficient flow to operate at its most efficient condition. The large turbocharger is essentially free-spinning at this point and not providing much compression. As the engine speed and load increase, the exhaust flapper is opened allowing the larger turbocharger to become active so it can supply air for the high power condition. A check valve is provided in the intake pipe so that the high air flow does not all have to pass through the small turbocharger compressor. This arrangement provides efficient operation over a wide range of engine operating conditions.

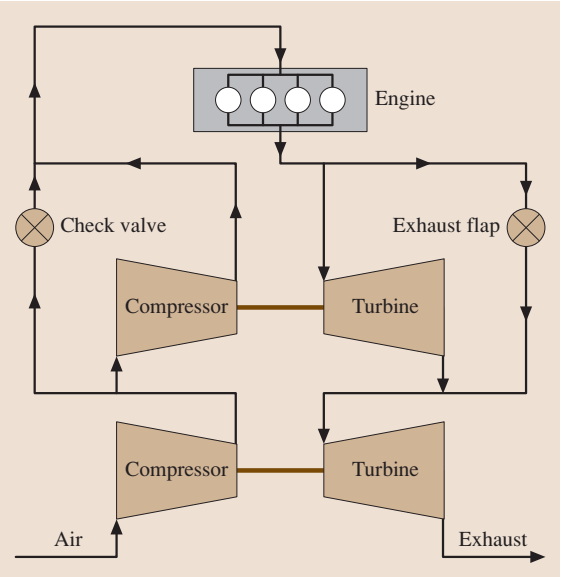


Fig. 10.76 Twin turbocharger configuration

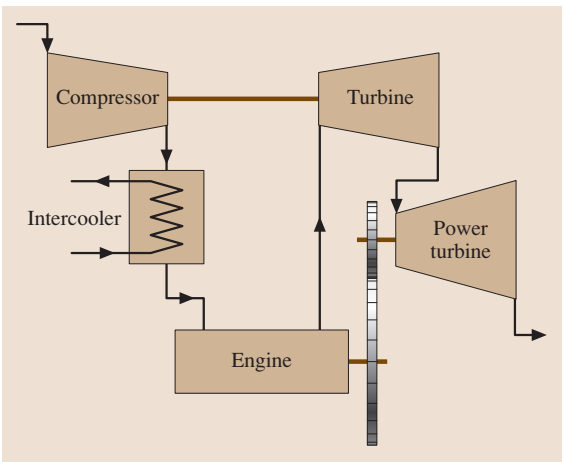


Fig. 10.75 A turbocompounded engine

Efficiency Definitions

The efficiency of a turbocharger is determined by the efficiency of its various elements. The efficiency of the compressor is calculated from the ratio of the work that would be required for a reversible adiabatic (isen-

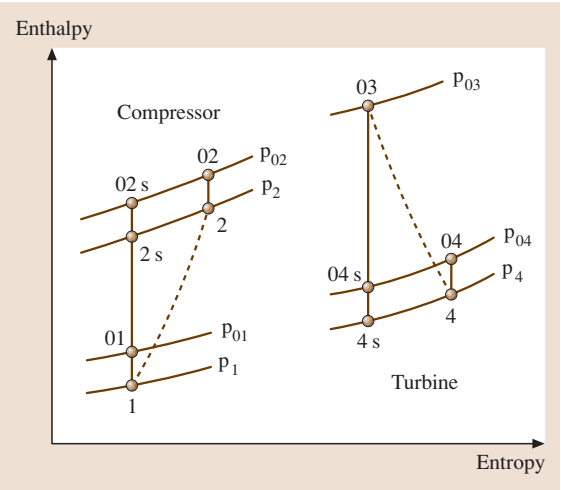


Fig. 10.77 State definitions for compressor and turbine efficiencies

tropic) process divided by the actual work for the process. The compressor efficiency can be calculated using stagnation states at the inlet and outlet, or since the kinetic energy leaving the compressor is not recovered, the efficiency can be calculated from the inlet stagnation state to the outlet static state. The latter choice is more conservative and allows the effectiveness of the compressor outlet diffuser to be included in the efficiency. The equation for compressor efficiency is provided below based on the state definitions given in Fig. 10.77

$$\eta_C = \frac{h_{2s} - h_{01}}{h_{02} - h_{01}} . \tag{10.79}$$

In a similar manner, the efficiency of the turbine from the inlet stagnation state to the outlet static state can be written as

$$\eta_T = \frac{h_{03} - h_{04}}{h_{03} - h_{4s}} . \tag{10.80}$$

Although the simple design of the turbocharger offers little opportunity for frictional losses, a mechanical efficiency can be defined that consists of the power delivered to the compressor divided by the power produced by the turbine

$$\eta_M = \frac{-\dot{W}_C}{\dot{W}_T} . \tag{10.81}$$

10.4.4 Fuel Systems

This section will cover gasoline and diesel fuel systems. The principles, main designs, key operating characteristics and controls of each system will be explained. Other important adjuncts such as low-pressure systems, filtration and sensors will also be covered.

Gasoline systems are divided into carburation and fuel injection; fuel injection is further broken down into throttle body, port injection and direct injection.

Diesel systems are divided into cam-driven and common-rail systems; cam driven systems are further divided into two main groups: pump-line-nozzle (inline, distributor, unit pump) and unit injector.

Gasoline Fuel Systems

Principle. Gasoline fuel systems can be divided into two main types: carburation and fuel injection. For all systems the goal is to achieve a stoichiometric air–fuel ratio, which is the ideal ratio whereby all of the fuel is completely mixed and burned. For normal gasoline this is usually around 14.7 : 1 air mass to fuel mass.

Carburation [10.21]. The simplest of all systems is the carburetor, which consists of the following subsystems:

- Inlet system to maintain a constant level of fuel in the reservoir
- Metering system to maintain the desired air–fuel ratio
- Accelerator-pump system to provide extra fuel during acceleration
- Power enrichment system to provide extra fuel during periods of high demand
- Choke system to provide a rich mixture for start and cold-engine operation

The carburetor uses the venturi principle: the inlet air flows through a necked-down area (venturi), where the flow increases in speed and decreases in pressure. A passage connects the fuel reservoir to the venturi; since the fuel in the reservoir is at atmospheric pressure, fuel flows from the reservoir to the lower-pressure area inside the venturi and then into the engine.

The pressure drop at the venturi increases with engine speed and with throttle position, thus causing fuel flow from the reservoir to the venturi to increase as engine speed and throttle position increase.

Fuel Injection [10.22]. Fuel injection can be divided into two basic types: manifold (throttle body and port) and gasoline direct injection (GDI). Manifold injection sys-

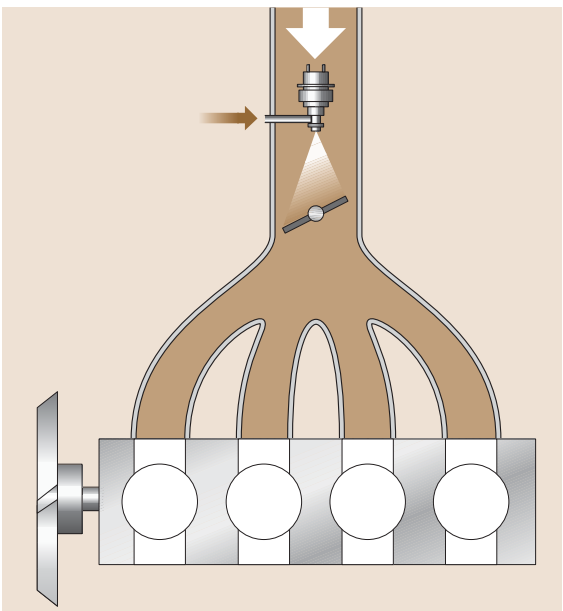


Fig. 10.78 Throttle body injection

tems allow only a homogeneous operating mode; GDI allows this and several other modes as well.

Throttle Body Injection. Throttle body fuel injection is also known as single point because there is a central point of injection: a single electromagnetically operated injector is located directly above the throttle valve Fig. 10.78.

Port Injection. Port injection is also known as multipoint because fuel is injected into every intake port, i. e., onto the cylinder's intake valve Fig. 10.79.

There are four types of port injection:

- Simultaneous fuel injection: all injectors open and close together. Half of the fuel quantity is injected in one engine revolution; the remaining half is injected in the next revolution.
- Group fuel injection: the injectors are combined into two groups. All injectors in a group open and close together. One injector group injects the total fuel quantity needed for its cylinders in one engine revolution, then the second set injects its total fuel quantity in the next revolution.
- Sequential fuel injection (SEFI): fuel is injected individually for each cylinder. Injectors are triggered is the same sequence as the firing order. Duration and start of injection (relative to each cylinder's top dead center) are the same for all cylinders.
- Cylinder-individual fuel injection (CIFI): the duration of injection (i. e., fuel quantity) can be varied for each individual cylinder.

Gasoline Direct Injection. Fuel is injected directly into each cylinder's combustion chamber Fig. 10.80.

An electric fuel pump delivers fuel to the high-pressure pump, which pressurizes the fuel to 50–120 bar (depending upon the engine operating condition) and sent on to the accumulator rail. Since all injectors are connected in parallel to the rail, they are all constantly pressurized and inject only when an electric signal is sent to each injector.

GDI allows not only homogeneous operation but also stratified charge, homogeneous stratified charge, homogeneous anti-knock and stratified-charge/catalyst heating.

Fuel Injection Control System. The injected fuel quantity is determined by the control system which consists of sensors that measure various input parameters, a processor to determine the optimum fuel quantity based upon the sensor input, and actuators which carry out the commands of the processor.

Examples of sensors: inlet air temperature, airflow, accelerator pedal position, throttle-valve angle, rail-pressure, Lambda, coolant temperature, etc. Examples of actuators: injectors, idle-air control valve, throttle valve, fuel-pressure regulator, etc.

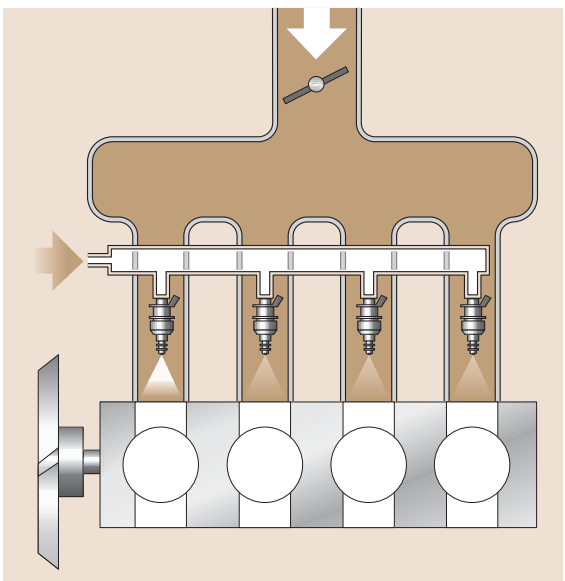


Fig. 10.79 Port injection

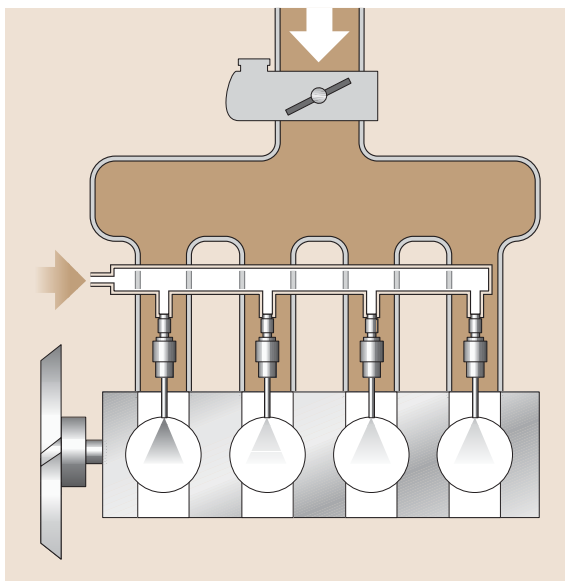


Fig. 10.80 Gasoline direct injection

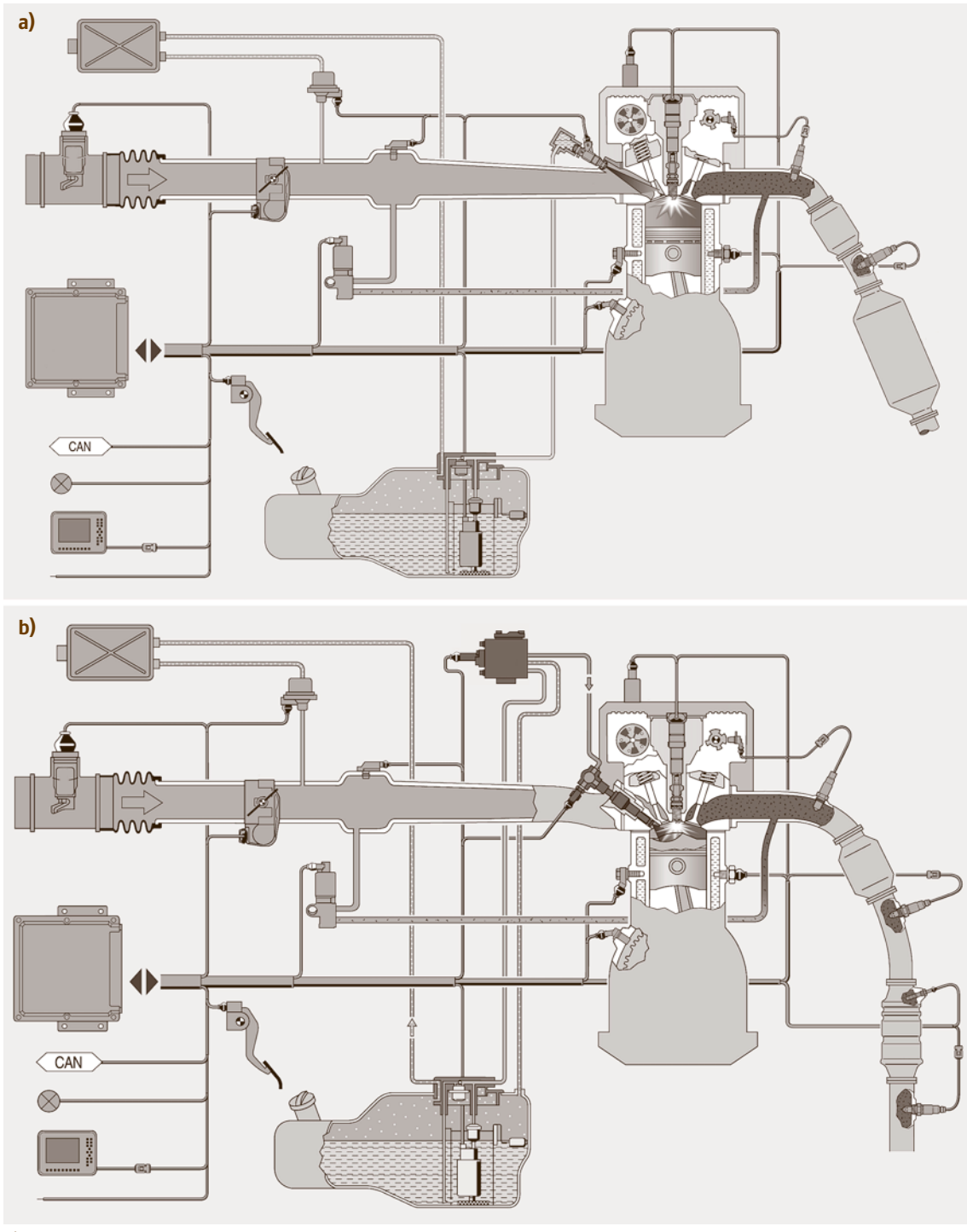


Fig. 10.81a,b Fuel injection control systems

The control system has to maintain the proper air–fuel ratio under the following engine operating modes:

- Start and warm-up
- Idle and part load
- Full load
- Acceleration and deceleration
- Overrun

The fuel injection control system may be combined with the ignition control system to allow for coordinated total engine management. The first image of Fig. 10.81 shows such a system for port injection and the second image of Fig. 10.81 shows a system for direct injection.

Fuel Supply System. For both carburetors and fuel injection, the fuel must be pumped from the storage tank by a mechanical or electrical pump, and must pass through a filter to remove impurities. The fuel pressure is regulated to a constant value.

Diesel Fuel Injection Systems

Principle. There are essentially two types of diesel fuel injection: indirect and direct. Since the indirect diesel injection (IDI) is for pre-chamber or whirl-chamber engines, both of which are seldom being applied today, we will concentrate on direct injection (DI).

In the DI process the fuel is injected directly into the highly compressed hot air in the combustion chamber above the piston (Fig. 10.82). A multi-hole nozzle is used to distribute the fuel uniformly in the combustion chamber to ensure rapid mixing. Very high injection pressures (up to 2000 bar) are required to properly and completely atomize the fuel.

For any combination of engine operating parameters, the fuel injection system must deliver the correct amount of fuel, at the correct time, at the correct injection pressure, with the correct timing pattern, and at the correct point in each cylinder's combustion chamber. Limits such as emissions, combustion pressure, exhaust temperature, engine speed and torque and vehicle-specific loads may need to be taken into account when determining the proper fuel injection.

Injection Characteristics. The key parameters and their corresponding units for every fuel injection system are:

- Injected fuel quantity (mm^3/stroke or mg/stroke)
- Injection pressure (bar or psi)

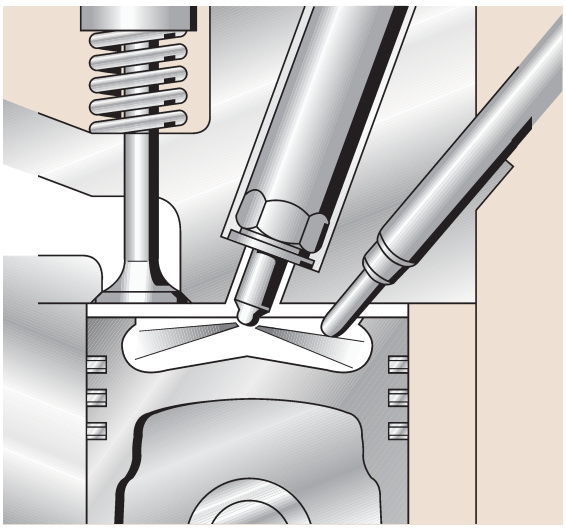


Fig. 10.82 Direct injection process

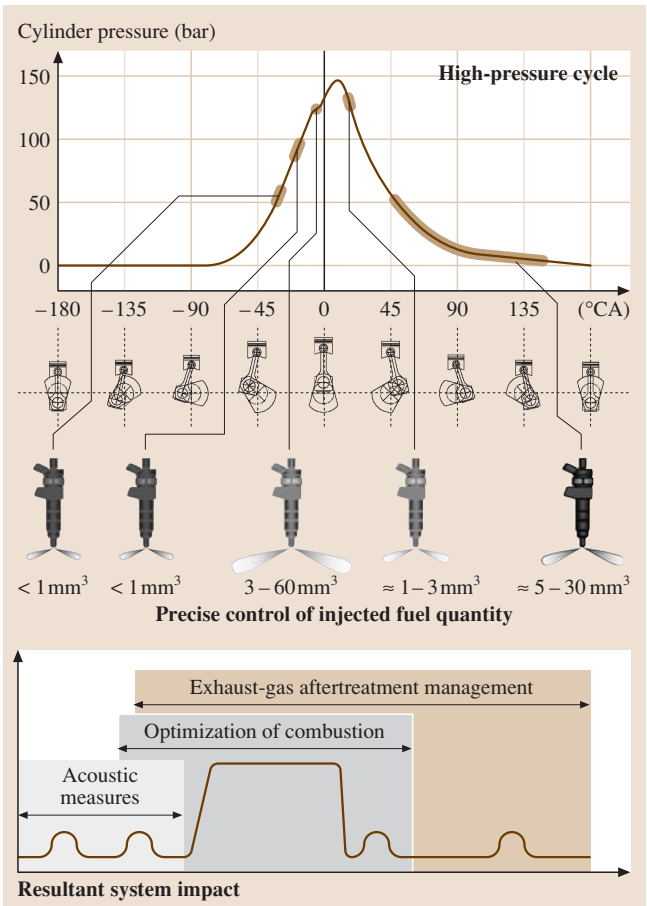


Fig. 10.83 Injection pattern

- Injection duration (degrees crank angle)
- Injection timing (degrees before or after the engine piston's TDC)
- Injection spray plume:
  - Number of plumes
  - The plume angle (degrees)
  - Location of the plume in the combustion chamber (height above piston bowl, position in cylinder)
  - The shape of the plume itself

Depending upon the type of fuel injection system, one or more additional injection characteristics may be available:

- Injection rate shape (mm^3 as a function of crank or cam angle)
- Injection pattern (number and form of injections during each combustion cycle, Fig. 10.83; up to five injections per combustion cycle may be required)
- Pre-injection 1: to reduce noise, improve warm-up (by avoiding misfire and white smoke)
- Pre-injection 2: to further reduce noise, improve warm-up (by avoiding misfire and white smoke)
- Main injection
- Post-injection 1 (close after the main injection): to reduce soot emissions
- Post-injection 2 (retarded): to act as a reducing agent for an after-treatment device

Fuel Injection System Designs [10.23]. To cover the wide variety of diesel engines in the marketplace (from motorcycles with 10 kW/cylinder up to railway locomotives with up to 1000 kW/cylinder), there is a corresponding variety of fuel injection designs. Figure 10.84 shows the various designs and Fig. 10.85 shows their corresponding injection pressure ranges. These can be divided into two main groups: cam driven and common rail.

Cam-Driven Systems. The cam-driven pumps can be further divided into two main groups: pump-line-nozzle and unit injector.

Pumping Arrangement of Cam-Driven Systems. The different designs within the pump-line-nozzle group:

- Inline fuel-injection pumps: one pumping unit per engine cylinder, mounted in a common housing
- Rotary fuel-injection pumps (axial-piston and radial-piston): one pumping unit for all engine cylinders; the fuel is distributed to each cylinder by a rotating shaft
- Individual-cylinder pumps: one pumping unit per engine cylinder, mounted separately and usually actuated by separate cam lobes on a common camshaft (Fig. 10.86)

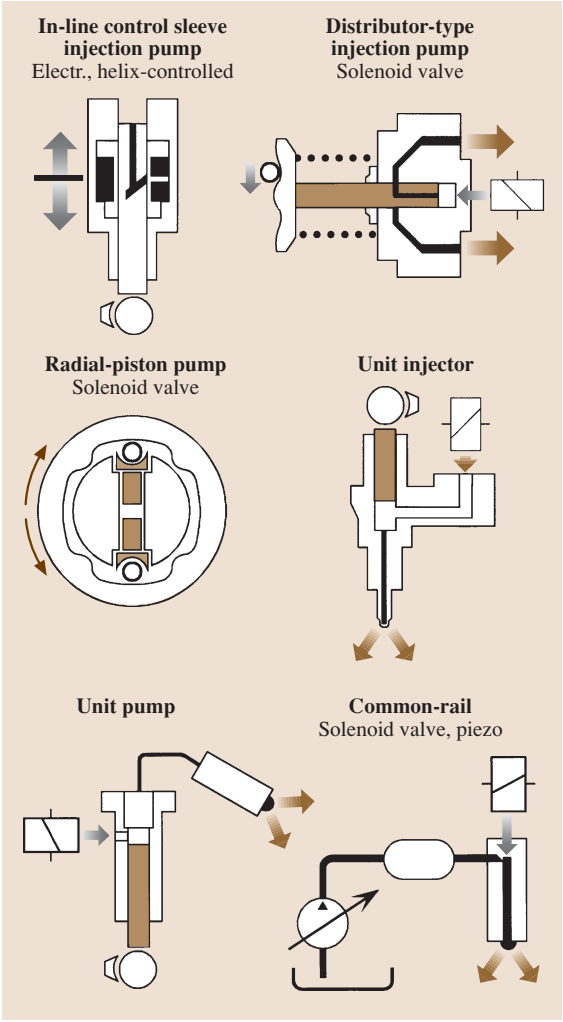


Fig. 10.84 Fuel injection systems

Each of the above has a high-pressure line to connect each pumping mechanism to its corresponding nozzle.

The unit injector combines the pumping unit and nozzle in one assembly for each engine cylinder; each unit injector is actuated by a separate cam lobe on a common camshaft (Fig. 10.87).

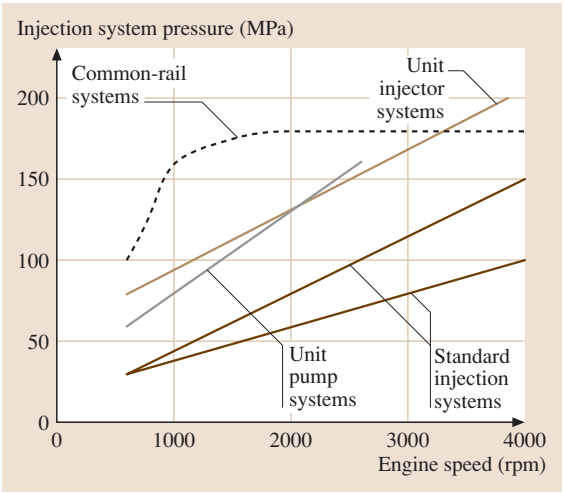


Fig. 10.85 Pressure ranges of fuel injection systems

Pumping Principle of Cam-Driven Systems. The pumping principle is essentially the same for all cam-driven systems: the cam lobe, which is driven by the engine's crankshaft and is thus phased to the crankshaft, pushes a plunger which pressurizes the fuel. This pressurized fuel travels to the nozzle and as soon as the pressure exceeds the nozzle's opening pressure, the nozzle needle lifts and fuel is injected into the engine.

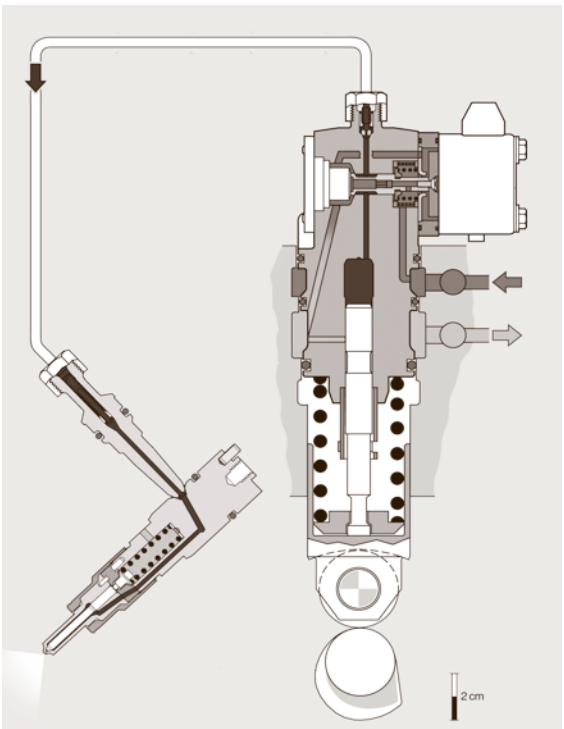


Fig. 10.86 Individual-cylinder pump

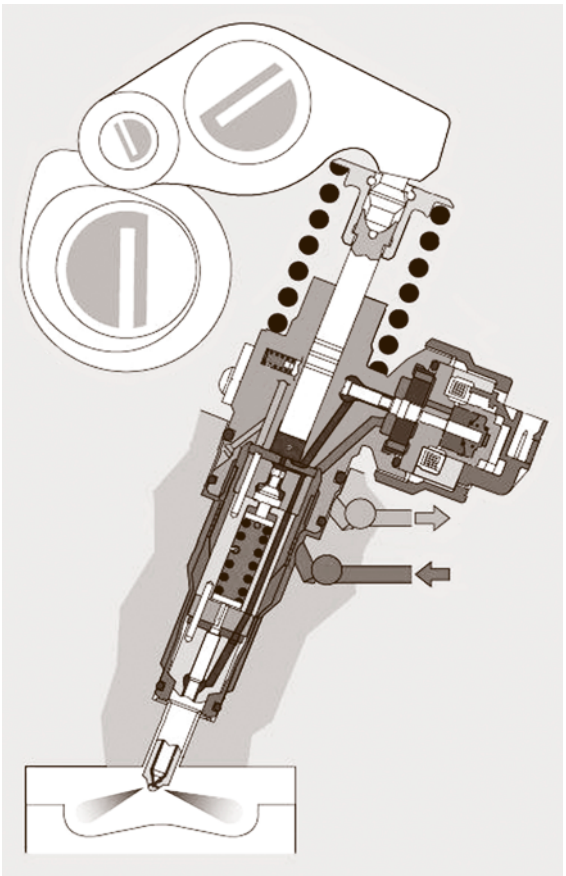


Fig. 10.87 Unit injector

Control System of Cam-Driven Systems. The fuel quantity is controlled by varying the length of delivery, which is accomplished by varying either the beginning of delivery (and keeping the end constant), the end of delivery (keeping the beginning of delivery constant), or by varying both. Governing can be mechanical, pneumatic, electromechanical, or electronic, whereby the majority of new applications are electronic.

In addition to fuel control the timing may also be controllable, either by changing the phasing of the pumping cam to the crankshaft or by changing the point at which pressure starts to build up on the pumping cam.

Injection-rate and pilot injection may also be controlled in some instances.

Common-Rail Systems. This system offers the greatest flexibility in the choice of fuel-injection parameters.

Pumping Arrangement. The common-rail (CR) system utilizes a single pump to pressurize the fuel which is delivered to an accumulator rail. One injector per engine cylinder is connected to this rail by means of a high-pressure line (Fig. 10.88).

Pumping Principle. Fuel is delivered by the low-pressure system to the high-pressure pump, where it is pressurized by the pumping plungers (arranged either radially or inline) and sent from there to the accumulator rail. Since all injectors are connected in parallel to the rail, they are all constantly pressurized and inject only when an electric signal is sent to each injector.

Control System. The start of injection (timing) and duration of injection (fueling) is controlled by a solenoid valve (electromagnetic or piezo) on each injector. When the actuating signal is sent to the injector, a coil or piezo stack inside the injector is energized, upsetting the pressure balance (on both sides of the nozzle needle) that was holding the nozzle closed. The needle then lifts and the injector delivers fuel to the engine. When the signal ceases, the pressure above the needle increases, forcing the needle closed and thus ending injection.

In addition, a sensor in the high-pressure circuit monitors the system pressure and sends this information to the electronic control unit (ECU) so that the pressure can be regulated to a value that is optimal for the engine's operating condition.

Low-Pressure System. Regardless of the fuel injection design, all fuel injection systems require a primary fuel pump to deliver the fuel from the fuel tank through the fuel filter to the injection pump.

Primary Fuel Pump. The primary fuel pump can be either mechanical or electric; the electric pump can be inside the fuel tank, mounted on the engine, or mounted on the vehicle.

Fuel Filter. The fuel filter must strain out impurities in the fuel, as contamination will cause wear, orifice plugging, component sticking, and seizures. The filter medium must be fine enough to trap small particles and the filter size must be large enough to assure adequate service life. Often a preliminary filter is used in addition to the main filter to extend the filter change interval.

In addition, filters should have the ability to separate water out of the fuel, as water will cause wear, corrosion, and seizures.

Many filters today have water separation, a water-in-fuel indicator, and a heater combined in one unit.

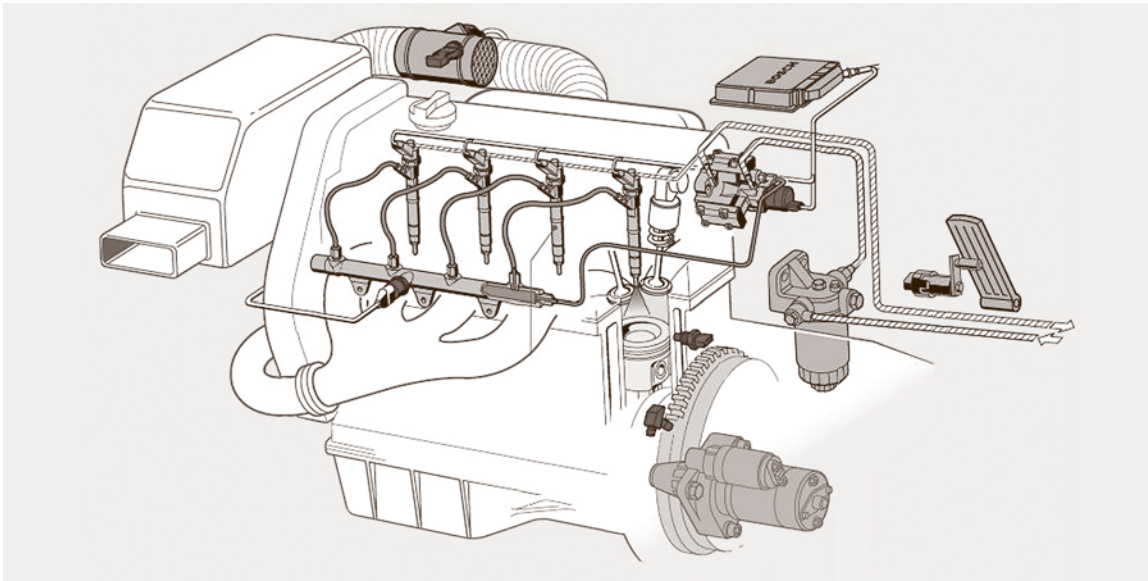


Fig. 10.88 Common-rail system

10.4.5 Ignition Systems

This section will cover the principles of ignition and the two main design types. It will explain how the high voltage needed for ignition is generated and the importance of ignition timing. Lastly it will cover spark plug design and function.

There are two basic types of ignition system: conventional coil and electronic. The conventional coil design can have one of three types of trigger; the electronic design can either have a distributor or be distributor-less. All designs use essentially the same method for generating high voltage, and all systems use the same design type of spark plug.

Principle

A gasoline internal combustion engine needs a spark to ignite the compressed air–fuel mixture in the combustion chamber. The spark is a discharge between the two electrodes that protrude into the combustion chamber. The ignition system generates the high voltage (up to 30 000 V) [10.24] needed to create the spark discharge and also initiates the spark to occur at the proper piston position (ignition timing).

Ignition System Design

Conventional Coil Ignition. This system consists of an ignition coil, ignition distributor, and spark plugs; see Fig. 10.89. As the coil is similar for all types of systems it is explained under point 3 (*high-voltage generation*).

The distributor rotates in sequence with the engine's crankshaft and at half of the crankshaft's speed for four-stroke engines or at crankshaft speed for two-stroke engines. One of three types of triggers is used in the distributor to control the current through the ignition coil:

- Mechanical breaker points: a mechanical switch that is closed and opened once per firing event by a cam located on the distributor shaft. The number of cam lobes equals the numbers of cylinders.
- Breaker-triggered transistorized ignition: this design is similar to mechanical breaker points except the primary ignition circuit is controlled by a transistor instead of by the breaker points. Only the control current is switched by the breaker points; this extends the breaker-point life and allows higher primary currents to be controlled.
- Transistorized ignition with Hall-effect trigger or induction-type pulse generator: the breaker points are totally eliminated and replaced by either a Hall-effect sensor or an induction-type pulse generator located in the distributor. The sensor or generator create one signal per cylinder; this signal is used to charge and discharge the coil.

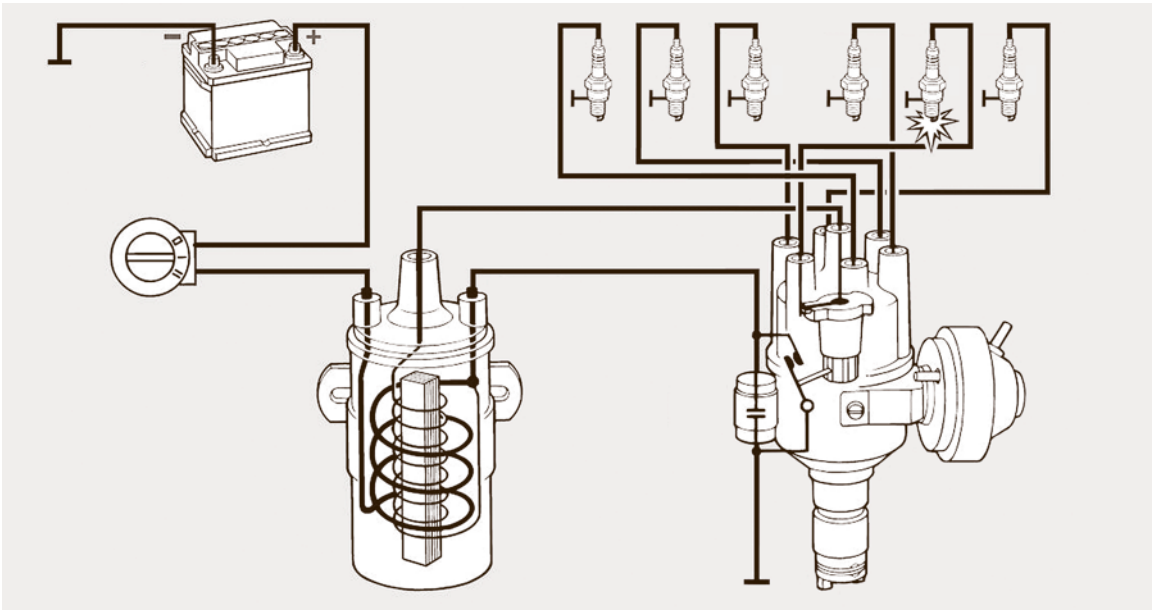


Fig. 10.89 Conventional coil ignition

A mechanical advance and vacuum unit define the proper ignition point as a function of engine speed and load.

Electronic Ignition. This system requires no centrifugal or vacuum-based timing control. Sensors monitor engine speed and load; these signals are sent to the ignition control unit, which determines the optimal ignition timing.

Engine speed is sensed by an inductive pulse sensor; engine load is sensed by a pressure (vacuum) sensor connected to the intake manifold. The ignition control unit uses these signals in a timing program map to determine optimal ignition timing for each speed/load point. Additional sensors may be used to monitor engine temperature, intake air temperature, throttle-plate position, and other operating parameters which are then taken into consideration when determining ignition timing.

This system may utilize a distributor to contain the engine speed/position sensor and to distribute the high voltage (Fig. 10.90a, or may instead be distributor-less (Fig. 10.90b. The latter system has a separate ignition coil for each cylinder or pair of cylinders. These coils may be mounted on the engine or directly attached to the spark plugs.

The ignition control system may be combined with the fuel injection control system to allow for coordinated total engine management (see Sect. 10.4.4).

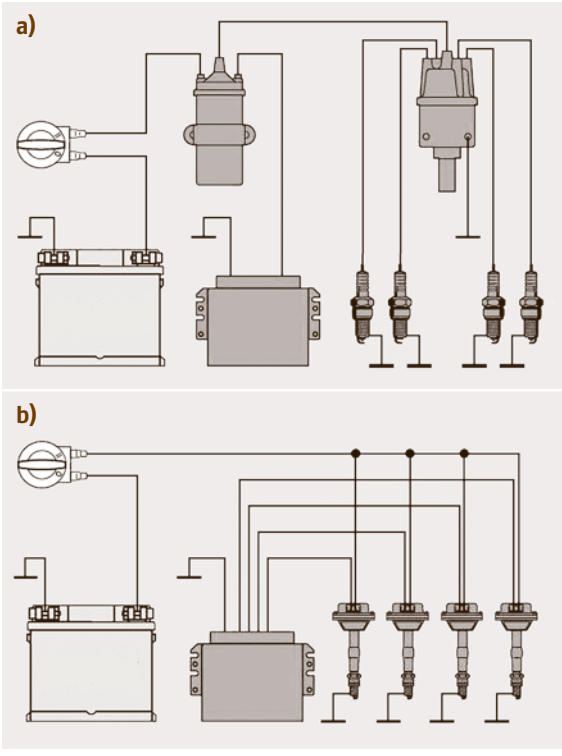


Fig. 10.90a,b Electronic ignition

High-Voltage Generation

The high voltage needed to bridge the spark-plug gap is typically generated by a coil. The coil consists of two copper windings (primary and secondary), an iron core, and a plastic casing. Energy is transferred from the primary winding to the secondary winding by means of magnetic induction. The current and voltage amplification from primary to secondary is the ratio of the number of coil windings.

Charging and discharging the coil to generate the high voltage needed for the spark plugs is essentially the same for all types of ignitions. The trigger closes and opens once for each cylinder's combustion event. When the trigger closes, current flows through the coil's primary winding and to ground. This produces a flux field in which ignition energy is stored. The time available for charging is determined by how long the trigger is closed (referred to as the dwell angle). The current is interrupted when the trigger opens.

The flux field in the primary winding produces high-tension voltage in the secondary winding. For a distributor-type system, this voltage is conducted to a center contact in the distributor cap. A rotor within the distributor rotates with engine rotation and distributes

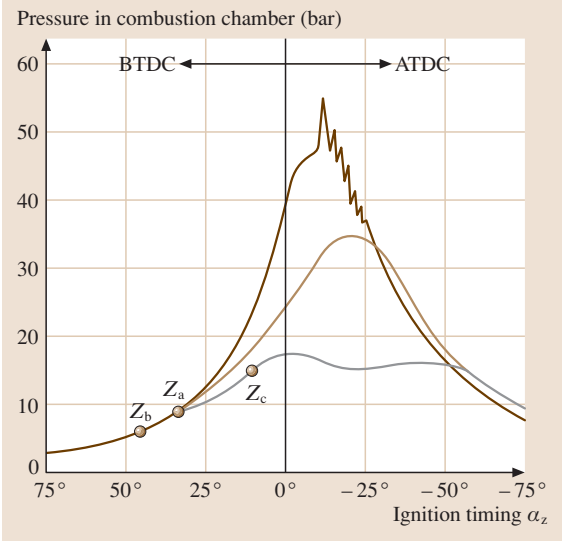


Fig. 10.91 Effect of ignition timing on combustion pressure

the high voltage to the spark plugs via the distributor cap and ignition wires. For a distributor-less system, the voltage is fed directly to the spark plug.

Ignition Timing

As engine speed increases, ignition timing must be advanced relative to piston top dead center to allow the air–fuel mixture more time to burn. As load increases, ignition timing must be retarded to prevent detonation (knock). The effect of ignition timing on combustion pressure is illustrated in Fig. 10.91.

Spark Plug Design

All ignition system designs include a spark plug, which consists of a:

- Terminal post which leads the current from the ignition wires to the center electrode
- Insulator which is made from a ceramic material and insulates the center electrode and terminal post from the shell
- Shell which houses the insulator and allows mounting to the cylinder head
- Gasket and seat which seals the combustion pressure
- Electrodes that form a gap that the high voltage must bridge to pass to ground, thus causing a spark

Electrodes may be made from a compound of corrosion-resistant nickel and copper, a composite with silver as a base, or platinum. The electrode gap determines the length of spark; the voltage required to jump the gap increases as gap width increases.

Figure 10.92 shows a spark plug cross section of two different electrode designs (front electrode and side electrode).

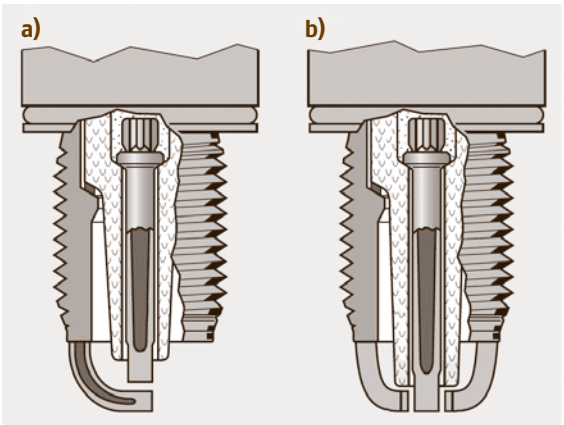


Fig. 10.92a,b Electrode designs

The heat range defines the operating temperature of the spark plug's insulator nose and should be chosen to maintain 500–900 °C. If the temperature is below 500 °C combustion residue will form on the insulator nose. If the temperature is above 900 °C the hot combustion residue gases promote electrode oxidation. Above 1100 °C auto-ignition may result. Figure 10.93 shows the proper operating range.

10.4.6 Mixture Formation and Combustion Processes

All fuel–air mixtures have limits for the mixture ratios that can be used in engines. When there is too little fuel, the mixture is said to be *lean* and flame propagation is slow and misfire is likely. When the fuel concentration is too high, the mixture is said to be *rich*, and the combustion produces products of incomplete combustion such as carbon monoxide.

Spark Ignition Engines

In spark-ignited engines, a homogeneous charge of fuel and air is introduced into the cylinder and a flame is initiated with a spark near the end of the compression

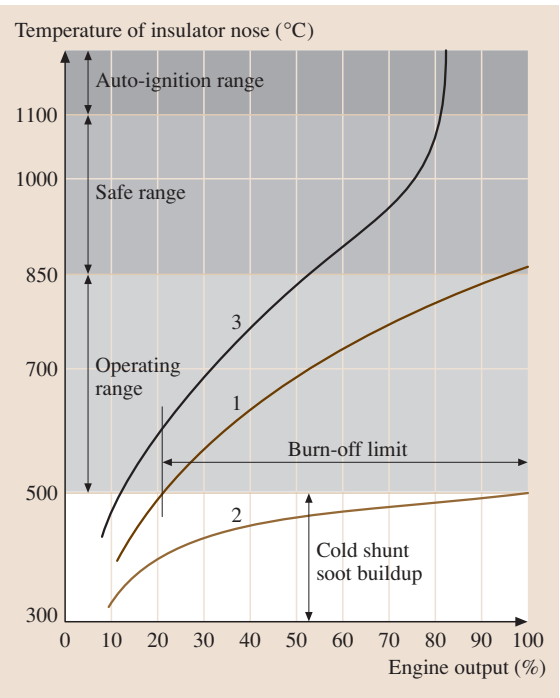


Fig. 10.93 Operating temperature of the spark plug's insulator nose

process. The flame propagates outward from the spark source. Turbulence is required for the flame to achieve the speed necessary to complete the combustion process before the exhaust valve opens. In the early part of the combustion process, immediately after ignition, the flame growth is slow. The combustion zone, called the flame kernel, is small and very little energy has been released. The progress of combustion at this point is sensitive to the balance of energy released by combustion and the heat lost to the cooler surroundings. Leaner mixtures and lower temperatures and pressures can slow the energy release rate and increase the likelihood of misfire. Cycle-to-cycle variations in this critical phase of the combustion process can cause large differences in the development of the cylinder pressure.

In some cases the combustion process may not proceed in the normal manner of a deflagration wave propagating away from the spark. In one such case, the gases in front of the flame front, which are compressed as the cylinder pressure rises from combustion of the gases behind the flame front, may spontaneously ignite. This *auto-ignition*, also called *knock*, can be extremely violent and significantly raises engine noise levels. If a large percentage of fuel is involved in the event, it can cause mechanical damage to the engine. Some fuels are more prone to auto-ignition than others. The resistance of the fuel to auto-ignition is characterized by the fuel's octane number. Fuels with high octane numbers can be operated with less chance of knock.

Basic thermodynamics indicates that engines with high compression ratio should have better fuel economy and performance. However, the higher temperatures and pressures that result from the higher compression ratio increase the tendency for knock. High octane fuels are needed for engines with elevated compression ratios. Because knocking combustion is so rapid, it tends to approximate the thermodynamic ideal of constant volume combustion. In fact, many engines demonstrate their highest efficiency when operated with light knock. However, as more fuel is consumed by the auto-ignition, the pressure oscillations in the cylinder tend to disrupt the thermal boundary layers in the cylinder and increase the heat loss from the combustion gases.

Another type of abnormal combustion is known as pre-ignition or surface ignition. In this case, the fuel–air mixture may be ignited by a hot surface in the cylinder. This might be a valve surface, a carbon deposit, or even a piece of head gasket that protrudes into the cylinder. Because the combustion may be earlier than the spark, the flame-caused pressure rise may occur during the compression process causing the cylinder pressure to become very high. It has been proposed that knock and pre-ignition can be coupled. Knocking can increase heat transfer rates and raise surface temperatures so that pre-ignition is more likely. The resulting pre-ignition raises the gas temperatures in the cylinder and makes knock more likely.

When the fuel–air mixture is too lean, the flame kernel grows slowly and there is an increased probability of misfire in the engine. The mixture ratio where the misfire becomes unacceptable is the lean operating limit of the engine. There is a corresponding rich operating limit where there is insufficient oxygen to sustain combustion. The chemically correct mixture is known as the *stoichiometric* mixture. The air–fuel ratio for maximum engine power falls on the rich side of the stoichiometric mixture and the best fuel economy tends to be on the lean side.

In order for a spark-ignited engine to operate continuously, the air–fuel mixture must be between the two flammability limits. However, exhaust emission concerns dictate that the mixture be held close to stoi-

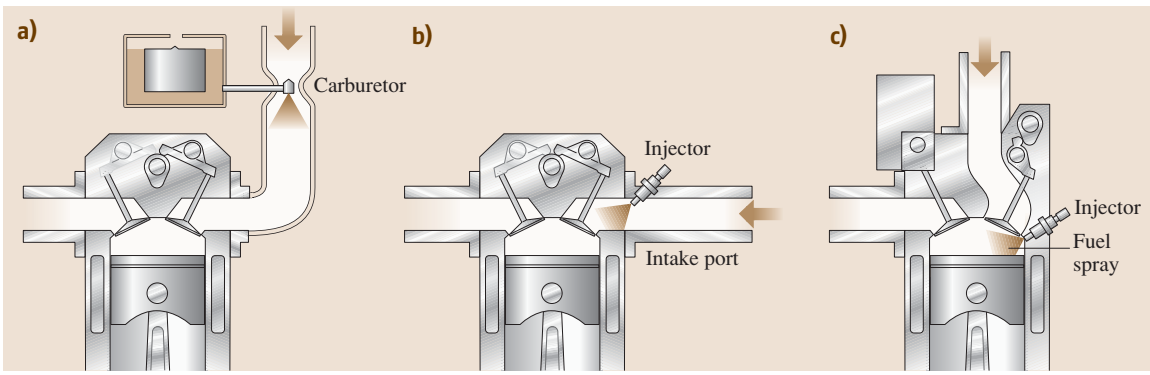


Fig. 10.94a–c Fuel–air mixture formation in spark ignited engines

chiometric. Except for cold starts and brief periods of high power demand, modern vehicles equipped with three-way catalysts and closed loop control of air–fuel ratio operate in a narrow band around stoichiometric.

Figure 10.94 shows three different approaches to fuel–air mixture formation in spark-ignited engines. The traditional approach of introducing the fuel with a carburetor into the intake air stream is only used for small engines and is obsolete for larger engines that are subject to emissions controls. Most engines today use port fuel injection, as shown in Fig. 10.94b. This approach provides very uniform air–fuel mixture between cylinders and excellent atomization of the fuel at all speeds. Figure 10.94c shows direct injection of the fuel into the cylinder. Although not yet widely used, this approach allows some degree of charge stratification in the cylinder and fuel economy that approaches the diesel engine.

Compression Ignition Engines

Compression ignition engines, also known as *diesel* engines, bring only air into the cylinder through the intake valve. The engines rely on compression of the air to produce sufficient temperature that the fuel auto-ignites soon after it is injected near the end of the compression process. In contrast to the homogeneous charge spark-ignited engine, the air–fuel mixture in the diesel engine is always heterogeneous. Since there is a distribution of fuel–air ratios ranging from very lean to very rich, there is always some location in the cylinder where conditions are optimum for auto-ignition and the cycle-to-cycle variability for diesel engines is very small.

Diesel engines run without throttles so they have the advantage of low pumping losses. Load is controlled by varying the amount of fuel that is injected into the cylinder. At light load and idle, the air–fuel ratio may be 70:1 or higher. At full load, the air–fuel ratio may be as low as 20:1. The stoichiometric ratio of 14.5–14.8 is generally not achieved because of high smoke levels.

Diesel engines can be categorized into direct injection (DI) and indirect injection (IDI), although indirect injection designs are now mostly obsolete. The IDI engines utilized a small chamber apart from the main chamber, called a prechamber or turbulence chamber, that was connected to the main chamber by a narrow passageway. Fuel was injected either into the separate chamber, or into the passageway, and the rapid air motion between the two chambers caused by the piston motion, provided excellent fuel–air mixing. This rapid mixing allowed high-speed operation of the engines but heat transfer and throttling losses exacted a severe fuel-economy penalty on IDI engines. Advances in fuel injection technology have allowed DI engines to operate at equivalent speeds with much better fuel economy. A DI engine with the fuel sprayed directly into a chamber in the cylinder formed by a toroidal recess in the piston is the most common configuration in modern diesel engines.

Alternative Combustion Systems

A major drawback of both spark ignited and compression ignited engines is their high NO_x emissions. Both engines require after-treatment to reduce NO_x to acceptable levels. Combustion systems that are homogeneous charge, like spark-ignited engines, but utilize auto-ignition, like a compression-ignited engine, have been developed. These engines utilize an air–fuel mixture that would ordinarily be quite lean, but by controlling the temperature, it can be made to auto-ignite in a gradual and controlled manner towards the end of the compression stroke. Temperature and air–fuel ratio can be optimized to reduce NO_x emissions to very low levels.

10.4.7 Fuels

This section will cover the basics of gasoline, diesel, and alternate fuels. It will start with the original of petrochemical fuels and the refining process. The basic composition and key characteristics of gasoline and diesel fuels will then be explained. Lastly an overview of fuel substitutes and alternate fuels will be given, whereby this cannot cover in detail the plethora of newly emerging fuels.

Petroleum Refining and Basic Organic Chemistry [10.25]

Most conventional fuels are made from petroleum crude oils, consisting primarily of paraffinic, naphthenic, and aromatic hydrocarbons. Raw crude oils have a wide range of densities ranging from as thin as water to as thick as tar. Crude oil is converted into usable products by means of refining; the most important products are gasoline, jet fuel, and diesel fuel. Other valuable products are heating oils, liquefied petroleum gas, lubricating oils and asphalt.

To convert crude oil the feedstock is typically distilled. Since the different components of crude oil (e.g., gasoline, diesel) have different boiling points, the lighter components (those with relatively low boiling points, e.g., propane and butane) rise to the top of the distillation column where they are drawn off. The next-heavier components (e.g., gasoline) are drawn off lower

on the column, then the subsequently heavier components (kerosene and then diesel) are drawn off towards the bottom.

The fuels must then be upgraded, usually by hydroprocessing (which uses hydrogen with a catalyst) to remove undesired components.

Fuels with higher boiling points are then cracked (broken down) into lower boiling points using very high temperatures and catalysts.

Gasoline

Basic Composition. Gasoline fuels for spark-ignition engines are hydrocarbon compounds, which sometimes contain oxygenous components to enhance performance.

Key Characteristics.

- Grade: usually stated as regular or premium; an indication of anti-knock property.
- Octane number: resistance to knock (pre-ignition).
- Density: weight per unit volume; energy content increases as density increases.
- Volatility: how easily the fuel vaporizes. The fuel must vaporize quickly for good cold starting but not so quickly as to cause vapor-lock. Volatility is characterized by the fuel's vapor pressure and/or evaporation points dependant upon temperature.
- Sulfur content: must be kept low to allow proper operation of the catalytic converter or other after-treatment device.
- Additives: may be used to enhance one or more of the properties stated above, or to protect against aging, contamination or corrosion.

Diesel

Basic Composition. Diesel fuels for compression-ignition engines are usually distilled from crude oil. They consist of a large number of different hydrocarbon compounds including *n*-paraffins, olefins, naphthenes and aromatic compounds. Diesel fuel ignites at $\approx 350\,^{\circ}\mathrm{C}$, much lower than gasoline, which ignites at $\approx 500\,^{\circ}\mathrm{C}$.

Key Characteristics.

- Grade: the standard to which the fuel must conform.
- Density: weight per unit volume; energy content increases as density increases.
- Viscosity: resistance to flow; low viscosity leads to leakage losses, while high viscosity may impair injection pump function.
- Cetane number: ease with which fuel ignites; combustibility increases as cetane number increases.
- Cold filter plugging point: the temperature at which the fuel clogs the filter.
- Flash point: the storage temperature at which flammable vapors are produced.
- Water content: amount of water in fuel; water causes corrosion and poor lubrication, leading to wear and seizures.
- Contaminants: foreign particles in fuel; the particles cause erosive and abrasive wear.
- Lubricity: measure of the fuel's lubrication properties; low lubricity causes wear and seizures.
- Sulfur content: amount of sulfur in fuel; sulfur does not harm the fuel injection system but will harm most after-treatment devices. The removal of sulfur by hydrogenation also removes the ionic fuel components that aid lubrication, reducing the lubricity properties of the fuel. Additives are thus needed to restore lubricity to a sufficient level.
- Oxidation stability: resistance to forming acids.
- Additives: may be used to enhance one or more of the properties stated above.

Alternate Fuels [10.26]

In addition to the standard fuels there are several alternatives to gasoline and diesel. These are often pursued in order to reduce emissions or to reduce consumption of nonrenewable fuels as most alternate fuels are from a renewable source.

It should be noted that alternate fuels are not always compatible with the fuel system (some require extensive modifications), and that they may increase one emission component while reducing another.

Alternate fuels can be divided basically into two categories: gasoline substitutes/additives and diesel substitutes/additives.

Gasoline Substitutes/Additives.

- Coal hydrogenation: coal and coke
- Liquefied petroleum gas (LPG): hydrocarbon mixtures (mostly propane) that are liquid at ambient temperatures under relatively low pressures
- Liquefied natural gas (LNG): methane that is cooled to $< -160\,^{\circ}\mathrm{C}$ and condensed to a liquid by compression
- Compressed natural gas (CNG): natural gas (mostly methane) compressed to high pressure
- Hydrogen: produced by electrolysis of water or from natural gas/coal

- Alcohol: methanol (usually made from natural gas but can also be made from biomass resources such as wood) and ethanol (made from grain or biomass). Either can be used alone or mixed with gasoline

Diesel Substitutes.

- Alcohol: methanol and ethanol as described above; usually mixed with diesel.
- Fatty acid methyl ester (FAME): often known as biodiesel, made from many sources the most popular of which are Rapeseed, soybean, sunflower and used cooking oils. Usually mixed with diesel in 5–50% concentrations, these bring lower PM emissions and improve lubricity but are critical regarding density, NO_x, water absorption, and stability.
- Diesel–water emulsions: reduce soot and NO_x but also lower power output
- dimethylether (DME): a gas-phase fuel, requiring extensive modification of the fuel injection equipment
- Oil sand/tar sand gasification
- Synthetic fuels: methane gas-to-liquid (GTL, tradename *Synfuel*), biomass-to-liquid (BTL, tradename *Sunfuel*), and coal-to-liquid (CTL).

10.4.8 Emissions

Emissions from engines are generally thought of as the harmful chemicals that are present in small amounts in the exhaust gas. However, vaporized fuel that is released during refueling or from leaks onboard the vehicle can also be a source of air pollution. Carbon dioxide, a major component of engine exhaust and an inevitable consequence of hydrocarbon combustion, was traditionally considered to be benign and not included as a hazardous air pollutant. Now that the role of carbon dioxide from fossil resources in global climate change has been identified, much more attention is being paid to this gas.

Emission regulations can be placed into two categories: those that limit emissions from specific engines or vehicles and those that limit the levels of emissions in specific spatial locations, without regard to their origin. The first category of regulations will be our primary focus as these are the regulations that engine and vehicle manufacturers are required to meet in order to sell their products. The second category of regulations tend to be the responsibility of cities or states, and while they may impose restrictions on the types of vehicles that can be sold within their jurisdiction, the sources of pollution are not limited to engines and vehicles.

Ambient Air Quality

Ozone has been the subject of what appear to be contradictory statements in the popular media. Ozone is described as essential to protect against skin cancer and is being depleted by chlorofluorocarbons (CFCs). At the same time, city dwellers are subject to warnings about

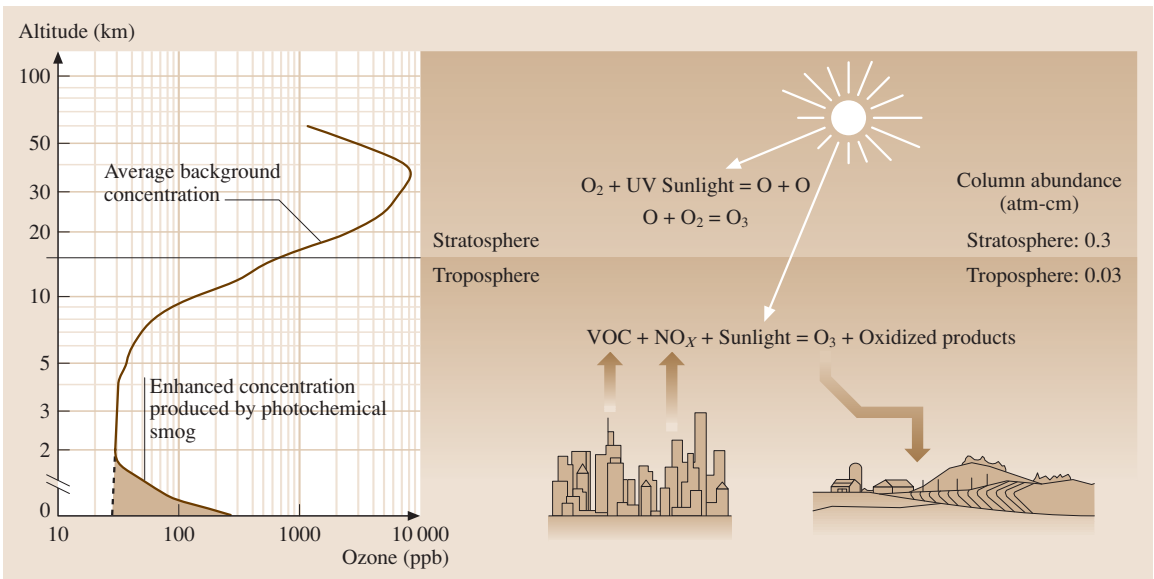


Fig. 10.95 Atmospheric ozone (after [10.27])

high ozone levels. The resolution of this apparent contradiction lies with understanding the role of altitude as explained in Fig. 10.95. In the stratosphere, above 15 000 feet, the concentration of ozone can be very high, approaching 10 000 ppb (parts per billion). This ozone is formed by reactions involving sunlight and oxygen. It filters the ultraviolet solar radiation that causes skin cancer. Some researchers have argued that this ozone is destroyed by reactions with chlorine atoms originating from CFCs.

At lower altitudes the natural concentration of ozone is up to 3 orders of magnitude lower than in the stratosphere. This is fortunate since high levels of ozone interfere with plant growth and are a strong irritant. High levels of ozone can be produced at lower elevations by reaction of volatile organic compounds (VOCs), carbon monoxide, oxides of nitrogen (NO_x), and sunlight. This complex set of chemical reactions produces a large number of different chemical compounds, many of which are harmful and irritating to people. Ozone, although only one chemical compound, is widely used as a measure of the overall concentration of the complex chemical mixture, sometimes known as *smog*. The EPA recently reduced the maximum allowable ozone concentration from 120 ppb averaged over a 1 h period to 80 ppb averaged over an 8 h period. Normal background levels of ozone are typically 20–40 ppb and can exceed 200 ppb during severe smog episodes [10.24].

Ozone is a secondary pollutant. It is not found in significant amounts in the exhaust of engines. However, compounds that are found in engine exhaust contribute to the formation of ozone, such as, VOCs, carbon monoxide, and NO_x. Another major source of VOCs is evaporative emissions. Evaporative emissions originate from losses of fuel during refueling as well as when the vehicle undergoes diurnal heating and cooling. These emissions are strongly tied to the Reid vapor pressure of the fuel, which is why it is now closely controlled in areas where air pollution is a problem. Evaporative emissions are only a problem with volatile fuels such as gasoline and its blends with alcohol. Diesel fuel's vapor pressure is so low that it does not contribute to evaporative emissions.

Regulated Pollutants

The Environmental Protection Agency (EPA) regulates the tailpipe emissions of both spark-ignition and compression-ignition engines in the United States. Regulated pollutants from spark-ignited engines include carbon monoxide, oxides of nitrogen, and unburned hydrocarbons. Compression ignited, or diesel, engines must meet requirements for particulates as well as these gaseous species.

Carbon monoxide is primarily determined by the engine's air–fuel ratio. When the engine is operated fuel-rich there is insufficient oxygen to convert all of the carbon to carbon dioxide, so a portion is converted to carbon monoxide. Carbon monoxide is actually an intermediate product in the oxidation of hydrocarbons and is always present in significant amounts during the combustion process. Measured levels in the exhaust are usually higher than would be expected because the oxidation of the carbon monoxide tends to be a slow process that is limited by the rate of the reaction of CO with the OH radical, as shown in the following equation

$$CO + OH \rightarrow CO_2 + H\,. \tag{10.82}$$

The concentration of the OH radical decreases rapidly as the in-cylinder temperature drops during expansion, leaving the CO frozen at an elevated level. An additional mechanism that affects homogeneous charge engines is the partial oxidation of trapped fuel that emerges from crevices or oil films during the expansion process when the temperature is too low to oxidize the fuel completely before the exhaust valve opens. In carbureted engines, rich-burning cylinders resulting from the nonuniform distribution of fuel between cylinders is an important source of carbon monoxide. Carbon monoxide emissions from diesel engines are generally well below regulation limits because diesels always operate with excess air.

Oxides of nitrogen (NO_x) consist primarily of nitric oxide (NO) and nitrogen dioxide (NO_2). Nitric oxide originates through three potential mechanisms that are usually categorized as fuel NO_x, prompt NO_x, and thermal NO_x. Fuel nitrogen can contribute to NO_x formation but is usually not important for engines because gasoline and diesel fuel contain small amounts of nitrogen. Prompt NO_x is formed by reactions between nitrogen and hydrocarbons during the combustion process. This mechanism also does not seem to be an important source of NO_x for engines. The primary source of engine NO_x emissions is thermal NO_x. This mechanism involves the following three reactions

$$O + N_2 \rightarrow NO + N\,, \tag{10.83}$$

$$N + O_2 \rightarrow NO + O\,, \tag{10.84}$$

$$N + OH \rightarrow NO + H\,. \tag{10.85}$$

Since these reactions require significant concentrations of the radicals O, N, and OH, they only occur at

high temperatures. The reactions also require significant time to equilibrate so most of the NO formation occurs in the post-flame gases. Virtually all NO_x control strategies, such as timing retard and exhaust gas recirculation (EGR) focus on reducing the temperature of the post flame gases.

Unburned hydrocarbon emissions from spark-ignited engines generally originate from fuel that is trapped in crevices, oil films, or deposits and is thus protected from combustion during the main combustion event. This sequestered fuel is released when the pressure drops during the expansion process but the temperature may be too low for complete combustion. Some of the fuel may burn to carbon monoxide but a significant portion will remain unburned or only partially burned and this will be released in the engine exhaust. Some of the products of partial combustion, such as olefins and aldehydes, are highly reactive and are strong contributors to photochemical smog reactions. Occasional misfiring cycles can also be a significant source of unburned hydrocarbon (UHC) from spark-ignited engines. UHC from diesel engines generally originate from fuel that has been overmixed with air so that the mixture is too lean to burn under the conditions in the cylinder. These conditions are most likely to be encountered at idle and light loads.

Regulated emissions from compression-ignited, or diesel, engines include the CO, NO_x, and UHC described for spark ignition (SI) engines, but also include particulates. Particulates from diesel engines are operationally defined as whatever collects on a filter when the exhaust is cooled to 52 °C after the filter has had a chance to equilibrate in a temperature and humidity controlled environment. The primary constituent is carbonaceous matter, usually referred to as soot, that originates from high temperature pyrolysis reactions in the fuel-rich regions of the cylinder. The carbonaceous particles provide sites for the condensation and adsorption of high molecular weight hydrocarbons as the combustion products cool and this portion of the particulate is often referred to as the soluble organic fraction (SOF) or the volatile organic fraction (VOF). These high molecular weight hydrocarbons may originate from the fuel but are more frequently associated with the lubricating oil. Particulate may also contain sulfates resulting from the reaction of fuel-based sulfur to sulfur trioxide and then to a variety of sulfate compounds which may be observed as small droplets of sulfuric acid. Finally, the particulate may include inorganic compounds resulting from engine wear and lubricant additives. Many of the compounds identified in the SOF are known carcinogens and the small size of the particulates (0.01–0.1 μm) increases the potential for their inhalation and retention in the lungs. For these reasons, the regulatory levels for particulate emissions have been progressively lowered so that after 2007 exhaust filtration technology has been required for most on-highway engines.

Measurement Instruments

A variety of instrumentation technologies have been developed to quantify the levels of pollutants in engine exhaust gases. Some techniques such as Fourier transform infrared spectral analysis have broad applicability and are widely used for engine development. For emissions certification, specialized instruments are still used that have been developed to measure specific species, and often over limited ranges. Oxides of nitrogen are measured with devices that take advantage of the chemiluminescent reaction that occurs when NO reacts with ozone to form NO_2 and oxygen. The photon of light that is emitted by this chemical reaction can be measured and directly related to the concentration of NO. Total NO_x can be measured by passing the exhaust gas through a catalyst that converts the NO_2 into NO before the gas is exposed to ozone.

Carbon monoxide and carbon dioxide are most commonly measured with nondispersive infrared (NDIR) absorption instruments that measure the amount of light of a specific wavelength that is absorbed by the exhaust gas. The wavelengths and path lengths for the light are chosen to provide the best sensitivity for the gas of interest. This technique requires that water vapor, a broad-band absorber of infrared radiation, be removed before the measurement can be performed. The water is usually removed by cooling the gas to condense the water or passing the gas stream through a chemical desiccant.

Unburned hydrocarbons are measured using a flame ionization detector. These devices contain a small hydrogen flame located between two electrically charged plates. A small amount of the exhaust stream is fed into the hydrogen flame and the hydrocarbon-based carbon atoms produce flame ionization that can be measured as an electric current between the charged plates. The particle filters and all connecting lines must be heated to prevent condensation of the hydrocarbon vapors before they enter the flame. The heated flame ionization detector (HFID) measures the number of carbon atoms associated with hydrocarbons in the exhaust and thus requires an assumption to be made regarding the chemical composition of the hydrocarbons. Measurements

performed on gasoline-fueled engines often assume the hydrocarbon has the same structure as hexane. Assuming the unburned hydrocarbons have the same chemical structure as the fuel is also a common assumption for both gasoline and diesel engines.

Particulates are measured by filtering a portion of the exhaust gas and then weighing the increase in mass of the filter. The temperature of the filter is carefully controlled because if it is too low an excessive amount of the unburned hydrocarbon vapors may condense on the filter. The techniques used to capture a representative sample of the exhaust gas and allow the determination of the amount of particulate during a transient test cycle are described in the following section.

Test Cycles

Emissions from passenger cars and other light-duty vehicles are measured while the vehicle is operated on a chassis dynamometer. The chassis dynamometer connects the drive wheels of the vehicle to a load absorber through a set of rollers. The device allows the vehicle to be operated in a controlled laboratory but with an accurate simulation of actual in-use driving conditions. Flywheels and dissipative absorbers are used to simulate vehicle inertia and air resistance so the vehicle can be operated over transient driving cycles. Test cycles that involve following a vehicle speed versus time curve that models different driving conditions are used for emissions certification. Trained drivers can follow these curves very exactly although most installations now use computerized throttle controllers.

The wide variety of transmission, drive line, and engine combinations used in heavy-duty applications precludes the emissions certification of vehicles. Heavy duty emissions testing focuses on testing the engine itself while it is outside the vehicle connected to a computer-controlled load absorber. The engine is operated over a 20 min cycle where both the engine speed and torque are specified on a second-by-second basis. The 20 min test cycle consists of four 5-min segments that model different types of city and highway driving.

Exhaust emissions for both chassis dynamometer and engine dynamometer test systems involve injecting some or all of the exhaust gas stream into a dilution tunnel to lower the temperature of the exhaust gas and to simulate the particulate agglomeration processes that occur when the exhaust enters the atmosphere. These systems are equipped with flow control systems so that a constant volumetric flow rate is maintained for the sum of the engine exhaust and the dilution air. A sample of the diluted exhaust gas is filtered and the weight of the particulate is measured. In the case of chassis dynamometers, the pollutant is expressed as g/mile and for the engine dynamometer it is expressed as g/horsepower-hour or g/kW h. The denominator of the engine dynamometer term is the total amount of work performed by the engine during the test cycle.

Gaseous emissions are also sampled from the dilution tunnel. They may be collected in special chemically inert bags to obtain an integrated total for the cycle, or measured second by second to investigate the effect of different parts of the test cycle on a pollutant of interest.

Both chassis and engine dynamometer testing are conducted under carefully controlled laboratory conditions. There have been some claims that this testing is not representative of emissions from in-use vehicles and attempts have been made to characterize in-use emissions using chassis dynamometers to test vehicles chosen at random from traffic flows. This experience has shown that significant numbers of vehicles have improperly operating emission control systems and actual emission levels are higher than would be expected from emission certification data. Measurements have been attempted using light absorption techniques from passing traffic but results have been mixed. Variability in measurement points, vehicle types, weather effects, etc. mean that extremely large numbers of measurements are required.

SI Engine Emissions Characteristics

The primary pollutants of regulatory concern for spark-ignited engines are carbon monoxide, oxides of nitrogen, and unburned hydrocarbons. The levels of these pollutants in the engine exhaust depend strongly on the engine's operating conditions such as spark timing, load, speed, and air–fuel ratio. While the engine speed and load are controlled by the operator (for a manual transmission), the timing and air–fuel ratio are set by the engine's electronic control module (ECM) to keep the levels of the exhaust pollutants within the range allowed by emissions regulations while ensuring adequate vehicle performance. Under cold starting and high-power-demand conditions the air–fuel ratio is calibrated to be fuel-rich. However, under all other operating conditions, the ECM maintains the air–fuel ratio close to the chemically correct, or stoichiometric, ratio.

The impact of timing, speed and most other operating parameters on carbon monoxide is secondary to the air–fuel ratio. There may be some effect on CO as these parameters are varied but this is most likely due to changes in air–fuel ratio or in causing the engine to operate at non-optimum conditions for oxidation of the CO.

As described above, oxides of nitrogen from spark-ignited engines are almost entirely thermally based and the two primary parameters that influence combustion temperature are spark timing and air–fuel ratio. Over the normal range of variation of spark timing used in modern engines, earlier timing (advanced) always increases emissions of oxides of nitrogen (NO_x) and later timing (retarded) always decreases NO_x. Since the spark timing that provides best fuel economy is usually well advanced from the timing needed to keep NO_x at a tolerable level, engine designers are confronted by a trade-off between the desire to provide good fuel economy while still keeping NO_x emissions at a level that can be controlled by the catalytic converter.

While spark-ignited engine UHC emissions are primarily dependent on design factors such as piston ring position, which controls the size of the crevice volume, they can be affected by speed, timing, load, and fuel–air ratio as these will affect the amount of fuel that is sequestered in crevices and the tendency to misfire. As mentioned earlier, evaporative emissions are another source of UHC from spark-ignited engines. While these can be controlled during engine operation by venting the fuel tank through a carbon canister to absorb fuel vapors, vehicle refueling still usually involves a release of vapors as capture systems are not required in most states. Evaporative emissions are best controlled by limiting the fuel's vapor pressure during warm weather.

In some parts of the United States, oxygenates are required to be added to gasoline to lower emissions. Ethanol is currently the most widely used fuel oxygenate. Methyl *t*-butyl ether (MTBE) is being rapidly phased out due to concerns about groundwater contamination resulting from fuel spills. Because of its lower oxygen requirement for combustion, ethanol decreases emissions of CO and UHC during cold starting and periods of high load. This is partially offset by higher evaporative emissions resulting from the higher vapor pressure when ethanol is blended with standard grades of unleaded gasoline.

Closed-Loop and Open-Loop Control. As described above, the pollutant species that need to be controlled from spark-ignited engines are carbon monoxide, unburned hydrocarbons, and oxides of nitrogen. It should be noted that the first two of these compounds need to have their oxidation process completed, while the third compound needs to be *un-oxidized* or reduced. This conflict in objectives is what makes the simultaneous elimination of the three compounds so difficult. Completion of the oxidation process can be accomplished by providing an excess of oxygen and then passing the exhaust gas through an oxidizing catalyst such as platinum. Reducing the oxides of nitrogen requires an oxygen-poor environment and an easily oxidized compound, called a reducing agent, to assist in breaking down the nitric oxide.

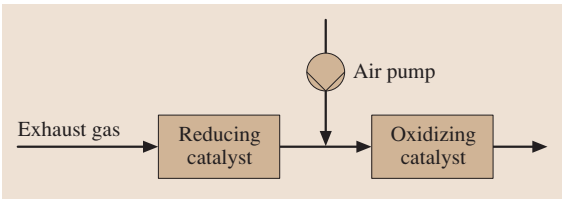


Fig. 10.96 Two-part catalytic converter

Early technology to achieve simultaneous control of the three pollutants required two operations, as shown in Fig. 10.96.

The engine was operated somewhat rich so that conditions were suitable for the exhaust entering the reducing catalyst, which eliminated most of the nitric oxide. Then an air pump supplied additional air to the exhaust stream to create the oxygen-rich conditions needed by the oxidation catalyst.

Modern catalysts, called three-way catalysts, can combine the oxidation and reduction functions into a single catalyst structure but require that the air–fuel ratio be controlled precisely in a narrow band around the stoichiometric value as shown in Fig. 10.97. Such exact

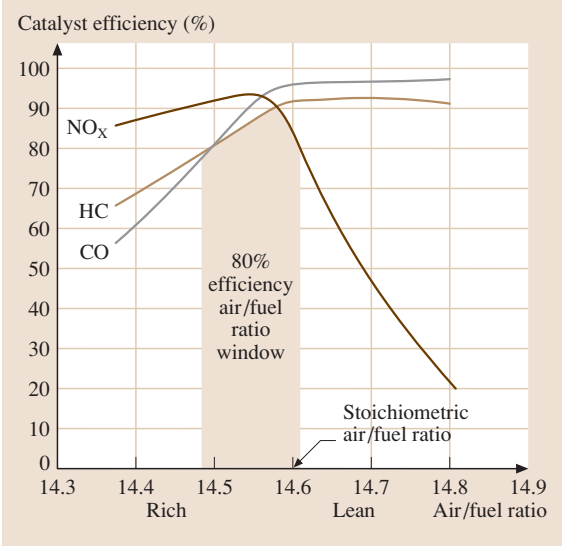


Fig. 10.97 Effect of air–fuel ratio on a three-way catalyst (after [10.28])

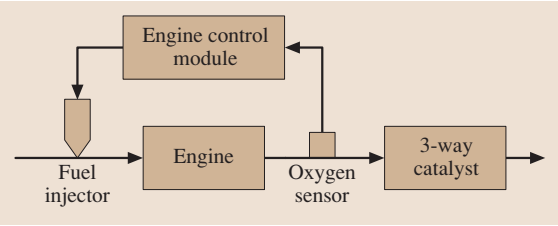


Fig. 10.98 Closed-loop engine control

control of air–fuel ratio could not be achieved without feedback from an exhaust oxygen sensor and closed-loop electronic control of fuel injection. Figure 10.98 shows the typical configuration of an oxygen sensor that measures the fuel–air ratio in the exhaust gas and then the engine's electronic control module adjusts the fuel injected to correspond to the air flow rate. This system maintains the air–fuel ratio entering the three-way catalyst in a narrow range around the stoichiometric ratio.

In reality, the nature of the oxygen sensor and the delay associated with the time required for changes in air–fuel ratio to reach the oxygen sensor results in an oscillation of the air–fuel ratio around the stoichiometric condition. This oscillation between rich and lean conditions enhances the catalyst's ability to provide the oxygen-poor conditions needed to reduce the oxides of nitrogen while using oxygen stored during the periods of lean operation to eliminate carbon monoxide and unburned hydrocarbons. Recent advances in oxygen sensor technology have incorporated heaters to decrease the time needed for the sensor to reach operating temperature and wideband sensing of air–fuel ratios to provide greater flexibility in control strategies over conventional rich-lean dual state sensors.

A complementary technique for NO_x reduction is exhaust gas recirculation (EGR). This technique directs a portion of the engine's exhaust gas back to the intake where it mixes with and dilutes the incoming charge. It reduces the flame temperature and the availability of oxygen without the fuel economy penalty that would accompany the equivalent NO_x reduction from spark timing retard. Exhaust gas recirculation is most effectively used under part-load conditions where it allows the throttle to be more open and thus can actually improve fuel economy by reducing throttling losses.

Compression Ignition (CI) Engine Emissions Characteristics

With the exception of carbon monoxide, exhaust emissions from compression ignited engines will be strongly affected by engine operating conditions. Carbon monoxide emissions are always low from diesel engines. They tend to be limited by factors such as late burning fuel that has insufficient time or temperature to combust completely. While more fuel is present at high load, the temperatures are lower at light load and there is a greater availability of fuel that has mixed beyond its lean combustion limit and thus can only react slowly.

Unburned hydrocarbon emissions from diesel engines are primarily a light load problem. At heavy loads the in-cylinder temperatures are high enough that the fuel readily burns to complete products. At light loads, fuel on the periphery of the fuel spray mixes with the large excess of air in the chamber and never reaches the temperature or air–fuel ratio needed for rapid combustion. This fuel will be emitted as unburned or partially burned hydrocarbons. Engines with the greatly retarded timing needed for NO_x control may also have difficulty keeping the fuel–air mixture in the piston bowl because the piston may be well down on its expansion stroke while fuel injection is still underway. This allows the fuel–air mixture to enter the crevice above the top compression ring and evade combustion in manner that is similar to spark-ignited engines.

As mentioned earlier, NO_x emissions are primarily dependent on in-cylinder temperatures. Since these temperatures are higher when the engine is operating at full load, the emissions of NO_x will be higher under these conditions. Diesel fuel injection timing is generally retarded to keep NO_x emissions low, in a manner that is similar to spark timing retard in spark-ignited engines. EGR can be used very effectively to reduce NO_x in diesel engines, especially if the EGR is cooled before it is mixed with the intake air. The cooling is usually accomplished using engine coolant so a portion of the exhaust energy is added to the engine's heat rejection load. EGR is usually accompanied by an increase in exhaust particulates.

Electronically controlled engines can vary the injection timing corresponding to speed, load, or other operating variables following detailed maps. This allows NO_x to be controlled with minimum impact on fuel economy and particulate emissions. NO_x is maximized when the fuel and air are mixed rapidly and combustion occurs near TDC. This is the condition that provides best conditions for minimizing soot production and maximizing its subsequent oxidation. These conflicting effects give rise to the phenomenon known as the NO_x-particulate tradeoff. Those measures taken to minimize NO_x (timing retard, EGR, lower swirl, etc.) tend to increase particulates and the converse is also true. Higher fuel injection pressures with smaller injec-

tor nozzle holes and lower intake air temperatures tend to move the NO_x–particulate tradeoff to more favorable operating points where both pollutants are reduced.

Exhaust gas recirculation can be problematic with diesel engines because the intake manifold pressure is frequently at a higher pressure than the exhaust manifold. Some systems direct the exhaust from upstream of the turbine to upstream of the compressor, which forces the exhaust gas to pass through the compressor. Other approaches use throttling to lower the pressure of the air entering the compressor so that exhaust can be drawn in from an ambient pressure source after the turbine.

Most diesel engines are equipped with turbochargers that allow demands for increased power to be met while still maintaining the air–fuel ratio at values that ensure low emissions. To meet NO_x emission standards, highly-rated engines use heat exchangers known as intercoolers or aftercoolers to reduce the temperature of the compressed air from the turbocharger compressor. Using ambient air as the exchange fluid for the intercoolers is the norm for most applications. Further improvements in engine air supply can be obtained from the use of variable geometry turbocharging. These turbochargers are equipped with variable area turbine nozzles so the exhaust velocity entering the turbine can be optimally matched to the engine's speed and load. This provides greater intake air boost pressures over a wider range of operating conditions.

Diesel engines cannot use the three-way catalyst technology that is used for spark-ignited engines because diesel engines always operate with excess air and it is difficult to reduce NO_x under lean conditions. Recent developments in catalyst technology have produced lean NO_x catalysts but their low efficiency has limited their acceptance. Catalytic systems that absorb the NO_x and then periodically release it as harmless gases when the engine is momentarily operated rich have been more successful and may be used in the near future.

10.4.9 Selected Examples of Combustion Engines

Compression Ignition Engine with Twin Turbo Technology (BMW 535d)

The world's first use of twin turbo technology for car diesel engines sets this engine apart from comparable engines. Along with increasing the specific power to 67 kW/l, most notably the speed range has been expanded to approximately 5000 min^{-1}. With a bore of 84 mm and a stroke of 90 mm, the engine design is based on the 3.0 l inline-six 530d. The engine's rated output is 200 kW and thus increased by 25% over its predecessor. Just like the power, the maximum torque was raised to 560 N m. The engine weight was increased by 14 kg, while specific consumption at maximum output was reduced by 7 g/kW h to 233 g/kW h compared to engines years older.

Crankcase and Transmission. The crankcase cast from pearlitic gray cast iron (GG25+) is based on the *deep skirt concept* already proven in preceding models. The side panels of the crankcase (crankcase skirts) are very deep. A special head design of the skirt area achieves more stiffness. The viscous damper first used in the 530d is used again in the 535d too. The damping effect is generated by varying shear forces in a highly viscose fluid in a narrow gap between the housing and swivel flywheel rim.

Mixture Formation and Combustion. The proven concept of BMW direct injection engines has also been adopted in this generation of engine. A central perpendicular injection nozzle and two intake and two exhaust ducts per cylinder are located on the cylinder head. One of the two intake ports is a swirl duct and the other a tangential duct. The two exhaust ducts are still combined in the cylinder head.

To reduce raw emissions, combustion has been concentrated in the outer zones of the piston combustion bowl. Moreover, the engine's compression ratio has been reduced from 17 : 1 (530d) to 16.5 : 1 by further optimizing the piston combustion bowl geometry.

Injection System. The injection system of the 535d is based on the second generation common rail system already used in the 530d with a maximum injection pressure of 1600 bar. The flow was elevated in the 530d by 20%. This injection system supports up to five injections with minimal injection intervals between injections per combustion cycle. The induction-side high-pressure pump's volume control generates maximum pressure as required. A micro-blind hole injection nozzle with six spray holes is used. Fuel consumption in the New European Driving Cycle (NEDC) test cycle is 8 l/100 km.

Exhaust System. The particle filter is an integral part of the exhaust system of the Euro-4 package. A second-generation filter with catalytically coated SiC substrate is used. With a 4.5 l volumetric capacity, the particle filter has a considerably longer service life than the first generation filter. Two exhaust temperature sensors and

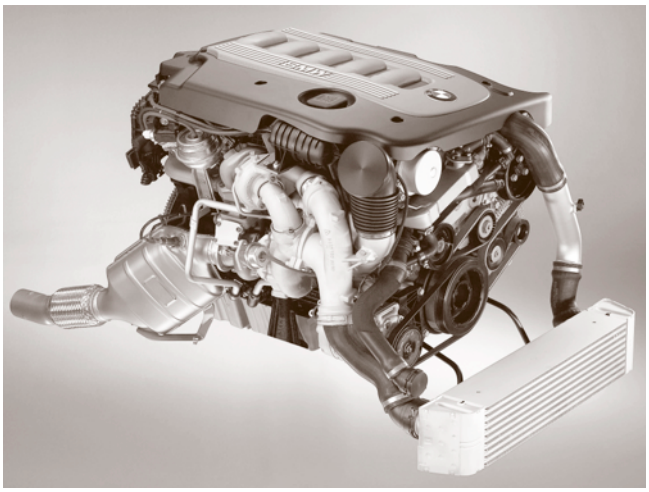


Fig. 10.99 BMW 535d

a pressure sensor are integrated in the exhaust gas line to monitor filter status. A temperature sensor integrated at the inlet of the upstream primary catalytic converter measures the exhaust gas temperature crucial for regenerating the particle filter.

Operating Principle of Twin Turbo Technology. Along with different supercharging concepts (multi-stage supercharging, a combination of mechanical and exhaust gas turbo charging) BMW favored the twin exhaust gas turbocharging after several tests.

This system consists of two differently sized turbochargers arranged in the induction and exhaust system branch as illustrated in Fig. 10.100. At lower speed ranges, the compressor bypass and the exhaust butterfly valve are closed as a result of which the entire exhaust mass flow is conducted through the small turbine. In this operating range, only the small turbocharger regulates the supercharging pressure. When the desired supercharging pressure is reached, the exhaust butterfly valve opens. The compressor bypass valve remains closed and part of the exhaust mass flow is conducted to the large turbine. The large compressor functions as a precompressor for the subsequent small compressor, which achieves supercharging pressures at average speeds (maximum supercharging pressure of 2850 mbar at 2500 min^{-1}).

From a certain speed onward, the small compressor can no longer generate additional supercharging pressure and throttles the induction mass flow. Depending on the load and upwards approximately 3000 min^{-1}, the compressor bypass and the exhaust butterfly valve open synchronously so that only the large turbocharger regulates the supercharging pressure supported by the waste gates.

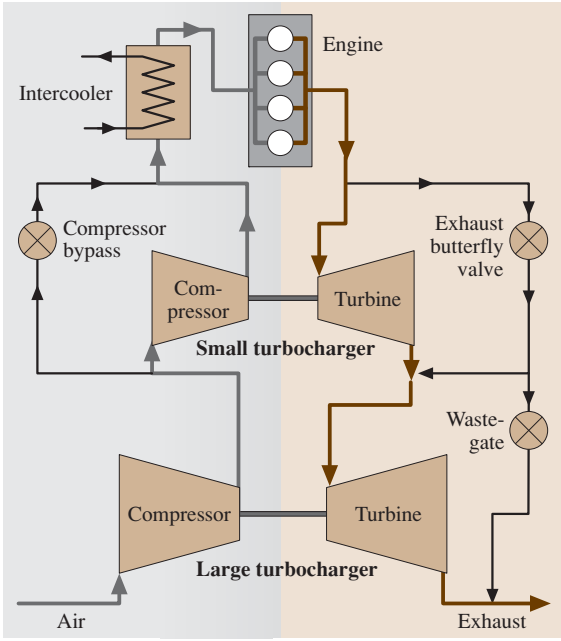


Fig. 10.100 Operating principle of twin turbo technology

Direct Injection Gasoline Engine with Downsizing-Concept and Dual Supercharging (VW 1.4l TSI)

This engine constitutes a logical contribution to downsizing modern gasoline engines. The direct injection FSI and dual supercharging used for the first time in this form achieve a power of 125 kW from only 1.4 l of displacement. Thus, the compressor, disengageable from the exhaust turbocharger, already reaches the maximum torque of 125 N m at an absolute supercharging pressure of 2.5 bar at low speeds. Along with increasing power, this downsizing concept most notably satisfies the requirement of low consumption of 7.2 l/100 km.

Basic Engine. The TSI unit is based on the four cylinder 1.4 l (66 kW) FSI engine used in the Golf V with a bore of 76.5 mm and a stroke of 75.6 mm and a compression ratio of 10 : 1. A basic reason for selecting this engine is the 1.4 l engine's modular design. Many modules could be carried over as a result of which the engineering was limited to a new cylinder crankcase and a water pump with an integrated magnetic clutch for engaging the mechanical supercharger.

The crankcase has an open deck (an open water jacket in the direction of the cylinder head) and deep skirt design (side panels far below the crankcase). Along with the advantage of simpler manufacturing, the open deck variant reduces cylinder barrel deformation when the cylinder head is bolted together. In order to withstand the high mean pressures of 21.7 bar in every operating situation, the material used is GJL (lamellar graphite cast iron), thus achieving a very low weight of 29 kg.

Transmission. Above all, great importance was attached to the engine acoustics. As opposed to the 1.4 l 66 kW, a steel crankshaft with 23% more stiffness was used for the TSI. This improves engine sound quality.

Calculation and development tools make it possible to develop a piston for use in a supercharged engine with a specific power of 90 kW/l. This light metal piston's combustion chamber bowl has a pronounce edge to control the flow. In order to provide the piston sufficient operational stability, oil ducts bolted into the main oil gallery inject with approximately 2 bar against the hot outlet side of the piston.

Finally, the piston pin diameter was enlarged because of the considerably higher ignition pressure.

Injection. The TSI engine is being used for the first time with a multiple hole, high-pressure injection valve with six fuel outlet bores. The nearly unlimited arrangement of the injection valve's spray makes it possible to form the fuel injection spray. Among other things, this not only optimally homogenizes the mixture but also prevents wetting of the intake valve when there is early injection. This reduces the hydrocarbons (HC) emissions.

The TSI injection pressure raised to 150 bar is generated by an adapted high-pressure pump. Compared to the FSI, its significant features include a longer cam stroke, the use of a roller tappet and the forged aluminum housing, all of which made it pos-

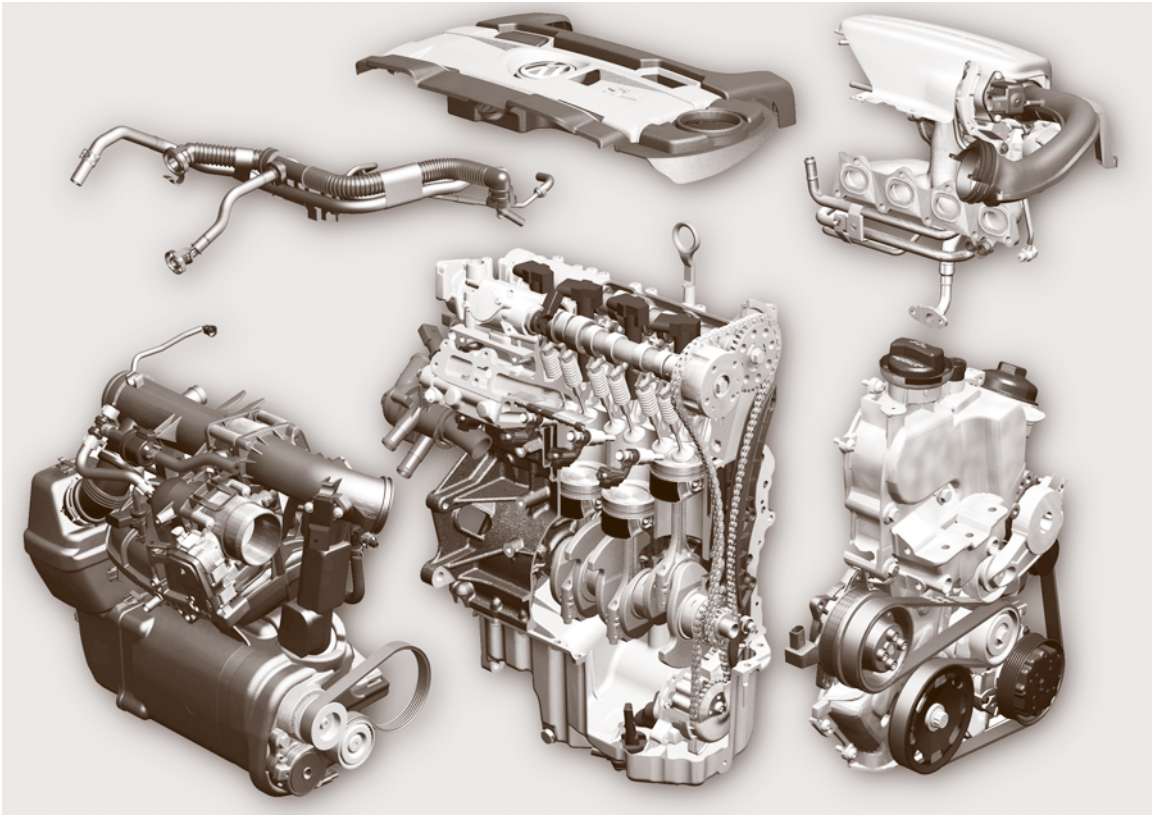


Fig. 10.101 VW 1.4l TSI

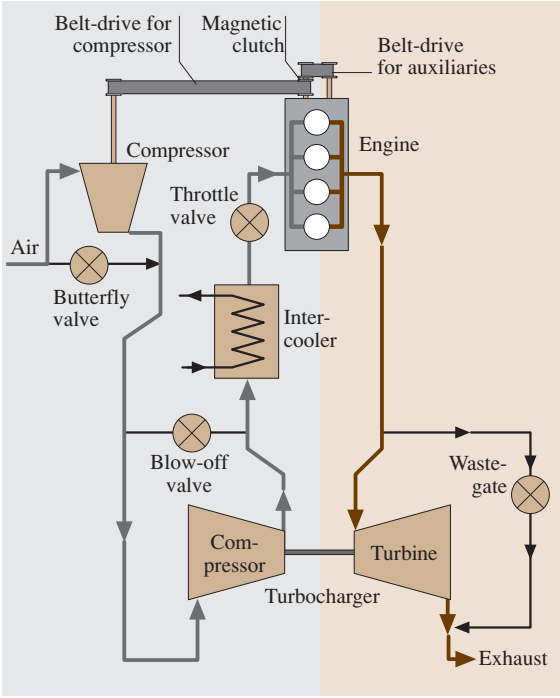


Fig. 10.102 Principle of twin supercharging

sible to approximately double the pump's mechanical stability.

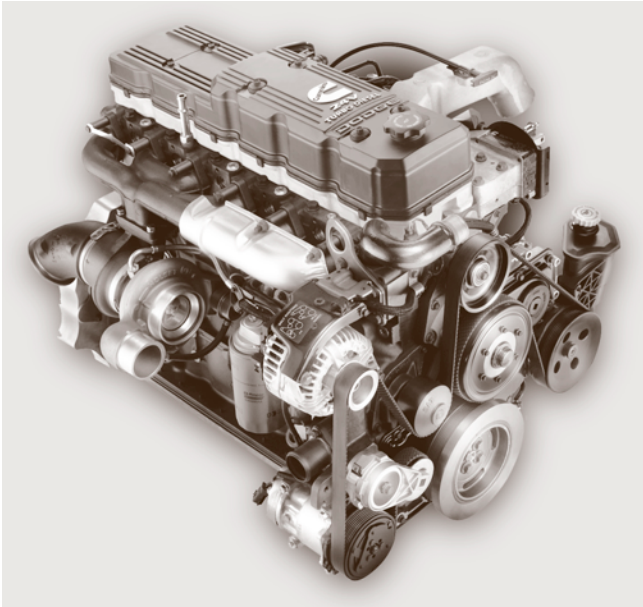


Fig. 10.103 Cummins 600 turbo diesel

Supercharging. Twin supercharging (see Fig. 10.102) basically consists of a Roots supercharger, an exhaust turbocharger and a butterfly valve.

Dependent on the engine map, the magnetic clutch on the water pump connects the compressor to the crankshaft. Inside the compressor is a countershaft transmission, which supplies a high torque above all when starting and in the low speed range. The butterfly valve enables a smooth transition between pure compressor and turbocharger operation.

By using the two charger units, maximum torque can already be produced at 1250–6000 min^{-1}. Since the exhaust gas turbocharger is designed for high efficiency, not enough boost pressure is available in the low speed range. Here, the compressor engages and bypasses the so-called *turbo lag*. At a speed of 3500 min^{-1}, the magnetic clutch disengages the compressor and the butterfly valve opens completely. From this point onward, the exhaust gas turbocharger alone produces the necessary boost pressure.

Diesel Engine for Heavy-Duty Pickups (Cummins 600 Turbo Diesel)

With 325 HP at 2900 min^{-1}, the Cummins 600 turbo is one of the most powerful engines available for the pickup truck market. Available as the engine in the Dodge Ram heavy-duty pickup, this diesel engine already generates a peak torque of 810 N m (600 ft-lbs) at 1600 min^{-1}. However, weighing 544 kg, the 600 Turbo's average consumption has dropped about 2% below that of the earlier model.

This engine consist of 30% fewer parts than comparable V8 engines in its performance class. Consequently, not only assembly time is cut but fewer repair costs are incurred because of its increased service life and durability.

Basic Engine and Transmission. Based on the B series, an agricultural machine engine sold over three million times, the Interact System B (ISB) was developed with an EGR radiator and a supercharging system. Predominantly used in medium-weight commercial vehicles, this engine series is the basic engine for the 600 Turbo diesel. A bore of 102 mm and a stroke of 120 mm produce a displacement of 5.9 l. This inline-six's compression ratio is 17.3 : 1. Push rods control the valves of the bottom-mounted camshaft.

The engine block is manufactured of cast steel and has a correspondingly high stiffness. Along with prolonging service life, this most notably reduces noise emissions.

The transmission is also designed to have a long service life. Not only the crankshaft optimized for weight and stiffness but also the forged fracture-split connection rods contribute to this.

These measures lead to considerably long running times so that an overhaul is only expected after an average of 350 000 miles.

Induction System and Cylinder Head. To enable an optimal turbocharge cycle, the Cummins 600 turbo diesel is equipped with four valves per cylinder. The redesigned intercooler as well as the turbocharger's enlarged compressor wheel and housing achieve an optimal air flow during the induction phase.

Reinforced inconel exhaust valves and cobalt stellite exhaust valve seats are used to increase the engine's service life.

Injection and Combustion. The 600 turbo diesel engine employs a Bosch high-pressure common-rail injection system. This system enables pilot injection before main injection. Thus the ignition delay time of the subsequent main injection is reduced considerably and combustion runs more gently, ultimately producing less combustion noise.

The injector is arranged centrally between the four valves and supports the target values for high efficiency and low emissions.

Exhaust System and Turbocharger. An *in-cylinder solution* and an oxidation catalyst reduce particulate and nitrogen oxide emissions considerably. A newly engineered piston combustion bowl likewise reduces pollutants.

Its compliance with the 2004 emission standards makes it possible to dispense with an external EGR line, which would add over 50 components to the engine's configuration and consequently make it more prone to failure. An expensive soot filter can be dispensed with for the same reasons.

A turbocharger with electronically controlled waste gate is integrated to further reduce emissions and reach the maximum power of 325 HP.

To spare the brake system when driving downhill, the Cummins 600 turbo diesel engine is equipped with an additional exhaust valve. It is closed as required and reduces the exhaust gas mass flow coming from the cylinder. This increases the in-cylinder pressure as a result of which the piston works against a stronger back pressure during the compression phase and crankshaft rotation is delayed.

Modern V8 Gasoline Engine with Variable Valve Timing

This engine is descended from the modular V8 and V10 engine family (MOD for short), developed by Ford in 1991. The engines are variably designed in terms of their cylinder heads (two-, three- or four-valve cylinder heads) and their use (trucks and cars). In conjunction with the six speed automatic transmission, electronically controlled throttle, variable valve timing and other state of the art engine technologies, it was possible to develop an engine that satisfies the requirement for greater power while simultaneously reducing gasoline consumption.

This engine is currently used in the Ford Explorer and a slightly modified version is used in the Ford Mustang.

Basic engine. The basic engine is a 4.6 l unit. The eight cylinders arranged in a V shape have a bore of 90.2 mm and a stroke of 90 mm. This engine also has a cylinder angle of 90° often used for V8 engines (compensation for higher-order inertial forces and moments).

Depending on its vehicle use, the cylinder block is made of aluminum (Ford Mustang) or cast iron (Ford Explorer).

Cylinder Head, Induction Pipe and Exhaust Manifold. In contrast to the central crankcase, the cylinder head in the Ford Explorer as well as the Mustang is made of

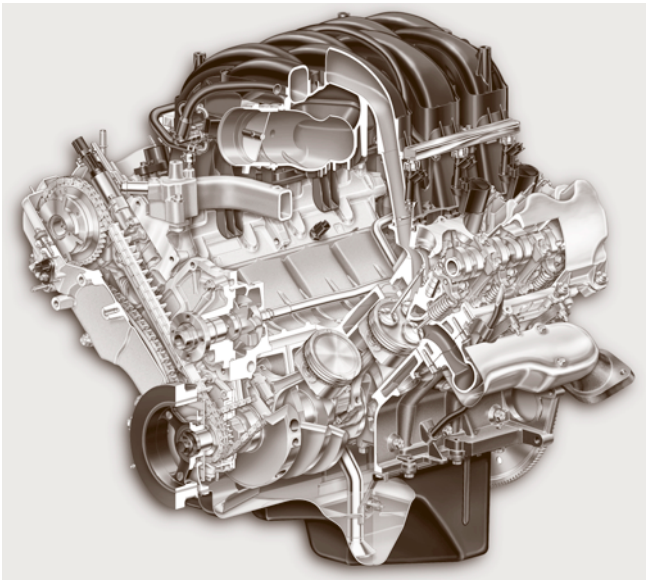


Fig. 10.104 Ford 4.6l single overhead camshaft (SOHC) 90° V8-engine (illustration courtesy of the Ford Motor Company)

aluminum. The three valve cylinder head is lighter and smaller than the four valve variant. The new cylinder head enables a higher compression ratio of 9.8 : 1 when 87 octane fuel is used.

The large dual intake ports create a direct path to the intake valves for enhanced flow behavior at high rpm. At low rpm and engine loads, a processor-controlled charge motion control valve (CMCV) in the induction line closes shortly behind the injection nozzle. This considerably increases the flow velocity in the induction tract as well as the in-cylinder flow resulting in a more ignitable and faster combustible mixture. Consequently, in conjunction with the variable valve timing, an optimal charge motion characteristic can be achieved in the induction tract. As a result, fuel consumption drops by 10% compared to the predecessor model.

In addition, the flow conditions in the longitudinally optimized intake manifold could be noticeably improved and a discharge of combusted gases from the cylinder accelerated.

Due to their mass inertia, extremely light intake and outlet valves make high engine speeds possible and simultaneously reduce fuel consumption through their enhanced frictional properties.

To minimize valve gear noise, the cam covers were made of magnesium.

Ignition System. The three-valve technology allowed arranging the spark plugs centrally in the cylinder head. This results in three advantages:

- The central position to the cylinder produces a symmetrical flame with complete fuel combustion. Since the proportion of uncombusted fuel is negligible, the engine can generate more power while simultaneously reducing emissions (uncombusted hydrocarbons).
- The narrow and more-oblong design of the spark plugs make it possible to enlarge the valve diameters. This results in better engine performance and lower fuel consumption.
- A new powertrain control module (PCM) controls the ignitions more precisely, which ultimately manifests itself in higher efficiency.

Variable Camshaft Timing (VCT). After two- and four-valve engines in the modular engine family have been put in the widest variety of vehicles, a 24 valve cylinder head in the V8 variable valve timing is being implemented for the first time in 2005.

A single overhead camshaft per cylinder bank and low profile roller-finger followers with low friction activate the intake and exhaust valves. The powertrain control module electromagnetically changes the oil flow for the hydraulic cam timing mechanism, which enables the camshaft to rotate opposite the drive sprockets. The mechanism can switch between fully advanced and fully retarded timing in only a few milliseconds.

VCT achieves an angular camshaft control of 50° CA. The *dual-equal* camshaft timing developed by Ford changes the intake and exhaust valve timing simultaneously. This system provides decisive advantages over fixed timing in the engine's complete speed range. Short seat timing at low speeds causes the cylinder pressure to drop less strongly as a result of which a high torque is generated. The slight valve overlap also reduces emissions. The seat timing increases at high speeds. The greater charge mass in the cylinder also increases engine performance.

This synchronous control of the timing enables constructing the cylinder head less complexly and with less weight than fully variable systems in which the intake valve is controlled separately from the exhaust valve.

References

10.1 K.-H. Küttner: *Kolbenmaschinen*, 6th edn. (Teubner, Stuttgart 1993), in German

10.2 H.T. Wagner, K.J. Fischer, J.D. von Frommann: *Strömungs- und Kolbenmaschinen*, 3rd edn. (Vieweg, Wiesbaden 1990), in German

10.3 H. Tschöke: *Vorlesungsskript Grundlagen Kolbenmaschinen* (Otto-von-Guericke-Universität, Magdeburg 2004), in German

10.4 K.-H. Grote, J. Feldhusen (Eds.): *Dubbel Taschenbuch für den Maschinenbau*, 21st edn. (Springer, Heidelberg 2005), in German

10.5 V. Küntscher: *Kraftfahrzeugmotoren – Auslegung und Konstruktion*, 3rd edn. (Technik, Berlin 1995), in German

10.6 H. Grohe: *Otto- und Dieselmotoren*, 10th edn. (Vogel, Würzburg 1992), in German

10.7 R. van Basshuysen, F. Schäfer: *Lexikon Motortechnik – Der Verbrennungsmotor von A–Z*, 1st edn. (Vieweg, Wiesbaden 2004), in German

10.8 KWW Crane GmbH: *DEPA-Druckluft-Membranpumpen* (CPFT, Düsseldorf 2005), in German

10.9 Ponndorf Gerätetechnik GmbH: *Hose Pumps* (Ponndorf, Kassel 2005)

10.10 G. Vetter: *Pumpen* (Vulkan, Essen 1992), in German

10.11 Alpha Laval Bran Lübbe GmbH: *Diaphragm Metering Pumps* (Bran Lübbe, Norderstedt 2005)

10.12 W. Hinze: *Kolbenpumpen und –verdichter, Bildsammlung zur Vorlesung* (Technische Hochschule Magdeburg, Institut für Kolbenmaschine und Maschinenlaboratorium (IKM), Magdeburg 1990), in German

10.13 R. Prager: *Technisches Handbuch Pumpen*, 7th edn. (Technik, Berlin 1987), in German

10.14 LEWA Herbert Ott GmbH: *ecofolow – Die innovativen Dosierpumpen* (LEWA, Leonberg 2005), in German

10.15 Eckerle Industrie Elektronik GmbH: *Internal Gear Pumps* (Eckerle, Malsch 2006)

10.16 R. Neumaier: *Rotierende Vedrängerpumpen* (Lederle Hermetic, Gundelfingen 1991), in German

10.17 Kräutler GmbH: *Screw Pumps Series M* (Kräutler, Lustenau 2006)

10.18 Netzsch-Mohno GmbH: *Nemo® -Pumpen* (Netzsch, Waldkraigburg 2006), in German

10.19 R. Bosch GmbH: *Mechanisches Benzineinspritzsystem mit Lambda-Regelung K-Jetronic* (Krebs GmbH, Stuttgart 1981), in German

10.20 Johnson Pumpen GmbH: *Impeller Pumps* (Johnson Pumpen, Löhne 2006)

10.21 M. Urich, B. Fisher: *Holley Carburetors and Manifolds* (HP, New York 1987)

10.22 R. Bosch GmbH: *Gasoline Engine Management Basics and Components* (Bosch, Stuttgart 2001)

10.23 R. Bosch GmbH: *Diesel-Engine Management*, 3rd edn. (Bosch, Stuttgart 2004)

10.24 R. Bosch GmbH: *Ignition Systems for Gasoline Engines* (Bosch, Stuttgart 2003)

10.25 R. Bosch GmbH: *Automotive Handbook*, 6th edn. (Society of Automotive Engineers, Warrendale 2004)

10.26 A Student's Guide to Alternate Fuel Vehicles (California Energy Commission, Sacramento 2006) (http://www.energyquest.ca.gov/transportation/index.html)

10.27 National Research Council: *Rethinking the Ozone Problem in Urban and Regional Air Pollution* (National Academy, Washington 1991)

10.28 J.T. Kummer: Catalysts for automobile emissions control, Prog. Energ. Combust. Sci. **6**, 177–199 (1980)

11. Pressure Vessels and Heat Exchangers

Ajay Mathur

This chapter is intended to present an overview of Pressure Vessels/Heat Exchangers and covers basic design concepts, Loadings & testing requirements relevant to these equipment. Design criteria, fabrication, testing & certification requirement of various Standards/Codes adopted in different countries are discussed on a comparitive basis to bring out similarities of features.

In order to complete the overview, a brief discussion is provided on commonly used Materials of construction and their welding practises along with updates on the on-going developments in this area.

The author is a Mechanical engineering graduatefrom M.S University-Baroda (India) & has over 20 years experience in design and fabrication of Pressure Vessels, exchangers, Skid mounted plants and Fired Heater modules for Refinery, petrochemical Nuclear & chemical plants in India & abroad.

11.1 Pressure Vessel – General Design Concepts

Pressure vessels are closed structures, commonly in the form of spheres, cylinders, cones, ellipsoids, toroids and/or their combinations and which contain liquid or gases under pressure. There are various other requirements such as end closures, openings for inlet/outlet pipes, internal/external attachments for support and other accessories.

11.1.1 Thin-Shell Pressure Vessel

When the thickness of the vessel is less than about one tenth of its mean radius, the vessel is called a thin-walled vessel and the associated stresses resulting from the contained pressure stress are called membrane stresses. The membrane stresses are assumed to be uni-

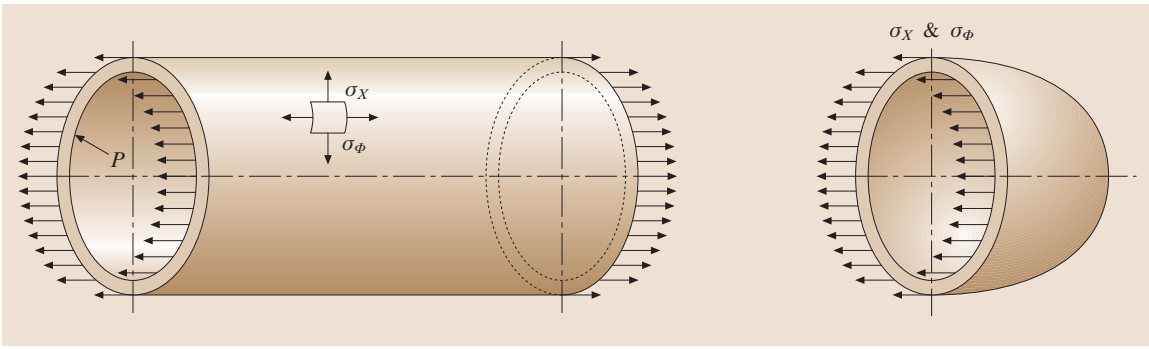


Fig. 11.1 Cylindrical shell with hemispherical heads

form across the vessel wall and act tangential to the surface.

Vessels Under Internal Pressure

The membrane stresses due to internal pressure are derived from application of equilibrium conditions to an appropriate element cut from the vessel shell, which in turn has been considered as a *symmetrically loaded shell of revolution*. Figure 11.1 shows the resulting membrane stress for a thin-walled cylindrical shell, both in the axial axis (called the *meridional* stress) and circumferential direction (also called the *hoop stress*), and the corresponding formulae are tabulated in Table 11.1. It can be seen that the thickness of the cylindrical vessel will be decided based on hoop stress, which is governing since its value is double than that of meridional stress.

In the case of a spherical shell, the circumferential stress is identical in all directions and its magnitude is half the hoop stress of a cylinder of the same radius.

Because the thin shell is assumed to offer no resistance to bending, radial stresses, if present, are small compared to membrane stresses and are generally ignored. This implies that the thin-shell analysis considers only biaxial stresses and therefore follows the *maximum stress theory*, wherein stress failure depends on the numerical magnitude of the principal stresses ignoring stresses in other directions.

Table 11.1 General vessel formula

Part	Stress formula	Thickness d	
		Inside radius	Outside radius
Shell			
Longitudinal direction	$\sigma_x = \frac{PR_m}{2d}$	$\frac{PR_i}{2SE+0.4P}$	$\frac{PR_0}{2SE+1.4P}$
Circumferential stress	$\sigma_\Phi = \frac{PR_m}{d}$	$\frac{PR_i}{SE-0.6P}$	$\frac{PR_0}{SE+0.4P}$
Heads			
Hemisphere longitudinal stress =circumflex stress	$\sigma_x = \sigma_\Phi = \frac{PR_m}{2d}$	$\frac{PR_i}{2SE-0.2P}$	$\frac{PR_0}{2SE+0.8P}$
Ellipsoidal		$\frac{PD_iK}{2SE-0.2P}$	$\frac{PD_0K}{2SE+2P(K-0.1)}$
2:1 semiellipsoidal		$\frac{PD_i}{2SE-0.2P}$	$\frac{PD_0}{2SE+1.8P}$
100%–6% torispherical		$\frac{0.885PL_i}{SE-0.1P}$	$\frac{0.885L_0}{SE+0.8P}$
Torispherical $L/r < 16.66$		$\frac{PL_iM}{2SE-0.2P}$	$\frac{PL_0M}{2SE+P(M-0.2)}$
Cone			
Longitudinal	$\sigma_x = \frac{PR_m}{2d\cos\alpha}$	$\frac{PD_i}{4\cos\alpha(SE+0.4p)}$	$\frac{PD_0}{4\cos\alpha(SE+1.4P)}$
Circumferential	$\sigma_\Phi = \frac{PR_m}{2d\cos\alpha}$	$\frac{PD_i}{2\cos\alpha(SE-0.6P)}$	$\frac{PD_0}{2\cos\alpha(SE+0.4P)}$

Vessels Under External Pressure

Thin-walled vessels under external pressure fail at stresses much lower than the yield strength due to the instability of the shell. In addition to the physical properties of the material of construction at the operating temperature, the principal factors governing the instability and the critical (collapsing) pressure P_c are geometrical, namely the unsupported shell length L, the shell thickness t and the outside diameter D_o. The solution to the theoretical elastic formula for the critical pressure at which the cylinders would collapse under external pressure depends on the number of lobes at collapse, which is a cumbersome exercise. To eliminate the dependency on the number of lobes, American Society of Mechanical Engineers (ASME) codes have adopted a simplified procedure using geometric charts developed for the critical stress ratio v/s L/D_o for different values of the D_o/t ratios. This critical stress ratio in turn is related to the collapsing ratio for a particular material in another chart, thus an estimate for the maximum allowable pressure may be determined for a given geometry. Several trials can be made to achieve an optimum solution for shell thickness, and maximum unsupported length (or stiffener spacing) for a given external pressure.

11.1.2 Thick-Walled Pressure Vessel

The membrane stresses used for evaluation of thin shells cannot be used for thick-walled vessels (with a thickness greater than their mean radius) subjected to internal pressure, since the radial stresses are significant and the stress distribution in the vessel walls varies across the thickness. The stress distribution across the thickness

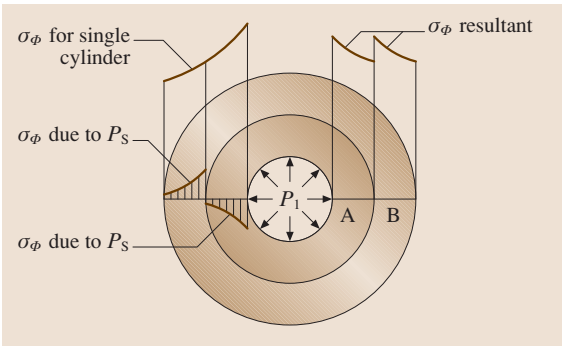


Fig. 11.3 Resultant stresses in a shrink-fitted multiwall vessel

for a vessel for internal and external pressure conditions are shown in Fig. 11.2a,b.

From Fig. 11.2, it is evident that the maximum circumferential stress occurs at the inside surface of the cylinder; however it is tensile in nature for internal pressure conditions and compressive in the case of external pressure. As per Lame's solution,

$$\sigma_\Phi = (+)\frac{PR_i^2}{R_o^2 - R_i^2}\left(1 + \frac{R_o^2}{R_i^2}\right), \tag{11.1}$$

$$\sigma_r = (-)\frac{PR_i^2}{R_o^2 - R_i^2}\left(1 - \frac{R_o^2}{R_i^2}\right), \tag{11.2}$$

the magnitude of stress for a thick-walled vessel is a function of the ratio of the outer to the inner radii.

Prestressing is a manufacturing technique by which compressive stresses are produced in the inside shell of a multilayer/multiwall vessel and/or tensile stresses are created on the outside shell of this vessel operating

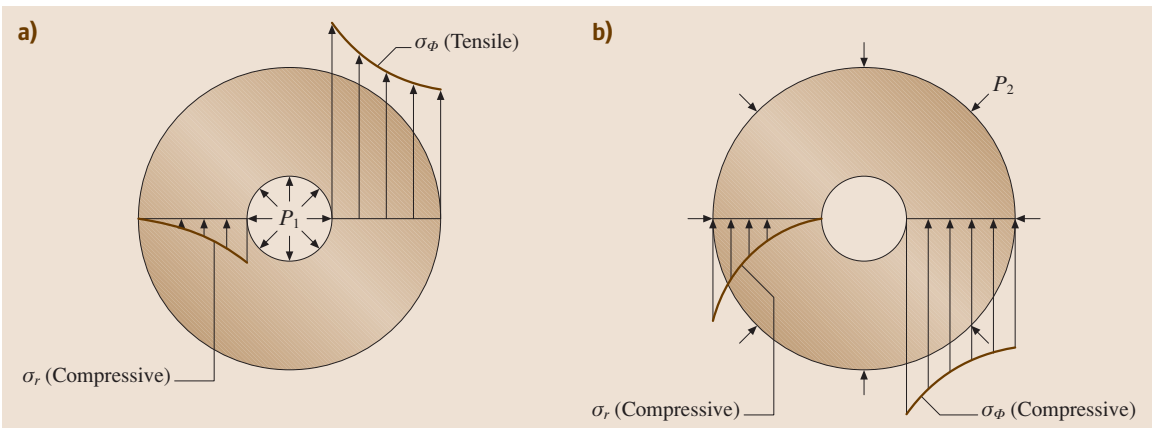


Fig. 11.2a,b Stress distribution for a thick shell: **(a)** internal pressure, **(b)** external pressure

under internal pressure. The prestresses so created help to neutralize the stress peaks in the existing stress distribution, making the material stress almost uniformly throughout the thickness, as shown in Fig. 11.3. This also considerably reduces the required wall thickness.

Prestressing is carried out by the following methods:

- shrink-fitting outer layers over the core, inner layer(s)
- wire/coil wrapping
- autofrettage

A detailed introduction to the basic theory of membrane stress and its application to commonly encountered elements of pressures vessels is presented in [11.1]. However, a brief discussion on other vessels components such as heads, nozzles and flanges is covered below.

11.1.3 Heads

There are three general categories of heads (also called dished ends):

Hemispherical: These dished ends are analyzed as the thin-walled spherical shells discussed earlier.

Ellipsoidal: Ellipsoidal heads are developed by the rotation of a semi-ellipse and have a 2:1 ratio of major R to minor axis h. These heads are the most frequently used end closures in vessel design, particularly for internal pressures greater than 10 bars and also for bottom heads of tall, slender columns.

Torispherical: Torispherical heads have a meridian formed of two circular arcs, a knuckle section with radius r, and a spherical crown segment with a crown radius of L. The maximum crown radius equals the inside diameter, which gives the same maximum membrane stress in the crown region as in the cylindrical region. The minimum knuckle radius is 6% of the crown radius, although a 10% knuckle is the most commonly used.

11.1.4 Conical Heads

A conical head is generated by the rotation of a straight line intersecting the axis of rotation at an angle, α, which is the half-apex angle of the formed cone.

The formulae for computing the thickness of different types of heads under both internal and external pressure are provided in Table 11.1.

There is a failure phenomenon in the knuckle region due to tangential stress under compression, which can occur through elastic buckling (circumferential wrinkles) at a stress much lower than the yield strength or through plastic buckling. A reliable analysis for predicting buckling failure has been introduced in the harmonized pressure vessel standard EN 13445, which will be discussed later.

11.1.5 Nozzles

Nozzles or openings are provided in pressure vessels to satisfy certain process requirements such as inlet or outlet connections, manholes, hand holes, vents and drains etc. These may be located on the shell or head according to the functional requirement and could be circular, elliptical or rectangular in shape. The basic construction of a nozzle connection consists of essentially short pieces of pipes welded to the vessel wall at an opening made in the wall. The other side would normally be a flanged end suitable for connection to the corresponding piping or to bolt on the blind cover (as in the case of a manhole). The complete nozzle may also be formed by rolling or forging to the required shape.

In addition to weakening of the vessel wall, the nozzle opening also causes discontinuity in the wall and creates stress concentration at the edges of the opening. This is compensated by providing reinforcement pads around the nozzle necks, which are suitably attached to the vessel wall. Rules are provided in every pressure vessel code to calculate the reinforcement requirement for all nozzles. At times, additional thickness is provided at the base of the nozzle wall itself; such nozzles are called self-reinforced nozzles.

11.1.6 Flanges

Tall columns are usually constructed in detachable sections for ease of fabrication, transportation, erection, assembly, and internal maintenance. Like the nozzles and the piping system, these sections (or heads) must be provided with end flanges with an arrangement for easy bolting and dismantling as required. A flanged joint therefore consists of a pair of flange; each is attached to one of the components to be joined and is held securely in place by a series of bolt or studs. A gasket is interposed between the two adjoining flange face. The joint must have structural integrity with (zero) min-

imum leakage during service. Several configurations of flanges in various construction materials and a large variety of gaskets are available; the selection is mainly dependent upon the service requirements.

The main consideration for the design of a flanged joint are:

1. to ensure a positive contact pressure at the gasket flange interface to prevent leakage in service. The gasket must be able to withstand the required sealing force,
2. the gasket sealing force is to be provided by bolt tightening without overstressing,
3. to ensure the structural integrity of flange sections and minimize flange deflections.

11.1.7 Loadings

Loadings or forces are the *causes* of stress in pressure vessels. It is important to identify areas *where* and *when* these forces are applied to pressure vessels. The stresses produced by these loads, which could be *general* or *local* are additive and define the overall state of stress in the vessel or its component. The combined stresses are then compared to the allowable stress defined by the pressure vessel code.

An outline of the various categories and types of loadings is summarized in Tables 11.2, 3, respectively.

11.1.8 External Local Loads

Stresses caused by external local loads at the points of attachment to the vessel must be assessed to keep these stresses within the prescribed limits. These loads are usually significant at nozzles, the vessel support region, at brackets, lifting lugs, and saddle supports for horizontal vessels. Since the contact area of the attachment is relatively small compared to the vessel area, a simplified form of the interface force distribution between the vessel and the support is assumed. The analysis is based on elastic stress analysis and stress categorization is used to compare the resultant stresses.

This approach is used in annexure G of the PD 5500 code and in the Welding Research Council (WRC) bulletins WRC 107 and 297 used while in the design of the ASME code. In 1991 the WRC published another bulletin, WRC 368, for the evaluation of shell and nozzle stresses due to internal pressure.

The analytical solution for evaluating localized stresses in the shell wall above the saddle supports for a horizontal vessel is based on the method developed by Zick and published in 1951.

The method used in the harmonized standard EN 13445-3 (clause 16) for evaluating nozzle loads is based on limit load analysis. This standard provides separate rules for calculating line loads for lifting eyes, supporting brackets, and saddle supports.

11.1.9 Fatigue Analysis

Vessels undergoing cyclic service and repeated loading may fail in fatigue due to progressive fracture of localized regions. The behavior of metals under fatigue conditions varies significantly from the normal stress–strain relationship. Damage accumulates dur-

Table 11.2 Categories of loadings

General loads	Local loads
General loads are applied more or less continuously across a vessel section	Local loads are due to reactions from supports, internals, attached piping and equipment such as mixers, platforms etc.
Examples	**Examples**
• Pressure loads – internal or external pressure (design, operating, hydrotest and hydrostatic head of liquid) • Moment loads – due to wind, seismic, erection and transportation • Compressive/tensile loads – due to dead weight, installed equipment, ladder, platform, insulation, piping and vessel contents • Thermal loads – skirt head attachment	• Radial load – inwards or outwards • Shear load – longitudinal or circumferential • Torsional load • Tangential load • Moment load – longitudinal or circumferential • Thermal loads

ing each cycle of loading and develops at localized regions of high stress until subsequent repetitions finally result in visible cracks, which grow, join, and spread.

Localized stresses at abrupt changes in sections such as at the head junction or a nozzle opening, misalignment, defects during fabrication, and thermal gradients are probable causes for fatigue failure. Special attention should be paid to manufacturing tolerances, nondestructive testing, and in-service inspection of vessels designed for fatigue.

All pressure vessel codes have established specific criteria for determining when a vessel must be designed for fatigue. Each code has adopted a methodology for carrying out fatigue analysis based on the use of fatigue curves obtained from test specimens. The fatigue design rules of some of the pressure vessel codes are discussed later.

11.2 Design of Tall Towers

11.2.1 Combination of Design Loads

The shell thickness of tall columns as computed based on internal or external pressure is not usually sufficient to withstand the combined stresses produced by the operating pressure plus weight, and wind or seismic loads. Combined stresses in the longitudinal direction σ_L due to pressure P, dead weight W and applied moment M, with W and M taken at the elevation under consideration, are calculated as follows:

1. On the windward side

$$\sigma_L = \left(\frac{PD}{4d}\right) + \left(\frac{4M}{\pi D^2 d}\right) - \left(\frac{W}{\pi Dd}\right), \quad (11.3)$$

$$d = \frac{\left[(PD/4) + (4M/\pi D^2) - (W/\pi D)\right]}{\text{allowable stress}}. \quad (11.4)$$

2. On the leeward side

$$\sigma_L = \left(\frac{PD}{4d}\right) - \left(\frac{4M}{\pi D^2 d}\right) - \left(\frac{W}{\pi Dd}\right). \quad (11.5)$$

For the design of a particular vessel, the value of the moment derived from either the wind or seismic loads is used in these expressions. This is due to the assumption that the wind and seismic loads are not expected to occur simultaneously and therefore the higher moment of the two is considered to be governing. These loads are based on site-specific data, which is obtained from one of the following national standards as applicable to the installation site.

Code for Wind Loads

- American Society of Civil Engineers (ASCE) 7, formerly American National Standards Institute (ANSI) A58.1
- Uniform building code (UBC)
- National building code (NBC)
- British standard (BS) 6399

Code for Seismic Loads

- ASCE-7
- UBC/NBC
- International building code (IBC) 2000
- Response spectrum data

Tall cylindrical vessels are normally designed to be self-supporting; they are supported on cylindrical or conical skirts with base rings resting on concrete foundations, firmly fixed to the foundations by anchor bolts. Detailed analytical methods for computing the thickness of the skirt components and sizing of the anchor bolts can be found in [11.2, 3], which also provide procedures for other types of supports such as lugs, legs, and rings due to wind and seismic loads.

11.2.2 Wind-Induced Deflection

A sustained wind pressure will cause tall columns to deflect with the wind. Most engineering specifications limit the deflection to 150 mm per 30 m of column height. The vessel is assumed to be a cantilever beam firmly fixed to the concrete pedestal and individual deflections induced by wind load and moments are calculated for each varying section of the vessel using the deflection formula for cantilever beams. The total deflection is then calculated using the method of superposition.

11.2.3 Wind-Induced Vibrations

Wind-induced vibrations can be caused by vortex shedding, the magnitude of which is dependent on wind velocity and vessel diameter. Vortex shedding results in severe oscillations, excessive deflections, structural damage and even failure. When the natural frequency

Table 11.3 Types of loadings

Steady loads	Non-Steady loads
These loads are long term and continuous	These loads are short term and variable
Examples	**Examples**
• Internal/external pressure • Dead weight • Vessel contents • Loading due to attached piping and equipment • Loading to and from vessel supports • Thermal loads • Wind loads	• Shop and field hydrotest • Earthquake • Erection • Transportation • Upset, emergency • Thermal loads • Startup, shutdown

of a column or stack coincides with the frequency of vortex shedding, the amplitude of vibrations is greatly magnified. After a vessel has been designed statically, it is necessary to determine if the vessel needs to be investigated dynamically for vibrations. Detailed methods for determination the need for dynamic analysis and a method for carrying out this dynamic analysis are provided in [11.3].

11.3 Testing Requirement

11.3.1 Nondestructive Testing (NDT)

NDT of the raw material, components, and finished vessel is important from the safety point of view. The most widely used methods of examination for plates, forgings, castings and welds are briefly described below.

Radiographic examination is done either by X-rays or gamma rays. The former has greater penetrating power, but the later is more portable. Single-wall or double-wall techniques are used for tubular components. Penetrameters, or image quality indicators, check the sensitivity of a radiographic technique to ensure that any defect will be visible.

Ultrasonic techniques use vibrations with a frequency between 0.5 and 20 MHz transmitted to the metal by a transducer. The instrument sends out a series of pulses, which are seen on a cathode ray screen after being reflected from the other end of the member. Reflections either from a crack or inclusion in the metal (or weld) can be detected on the screen; based on the magnitude and position of the signal, the location of the flaw can be ascertained.

Liquid penetrant examination involves wetting the surface with a colored fluid that penetrates open cracks. After wiping out excess fluid, the surface is coated with a developer fluid, which reveals the liquid that has penetrated the cracks. Another system uses a penetrant that becomes florescent under ultraviolet light.

Magnetic particle examination, which can only be used on magnetic material, is carried out by passing magnetic flux with the help of a probe through the part to be tested. Fine magnetic particles, which are dusted over the surface, tend to concentrate at the edge of the crack. To pick up all the cracks, the area is probed in two directions.

11.3.2 Destructive Testing of Welds

In contrast to the NDT methods, which are essentially predictive techniques, the mechanical integrity of welds is checked by testing sample test plates called *production test coupons*. The coupons are welded along with the actual vessel joint (usually a longitudinal seam) and thus are representative of the actual welding techniques employed for the vessel. The following tests are normally carried out on the test piece:

1. Tensile testing, which includes transverse tensile and all-weld tensile tests
2. Bend test – transverse bend/side bending
3. Macro-etching, hardness, impact testing
4. Intergranular corrosion test (IGC) for austenitic stainless-steel material/welds
5. Ferrite checking

11.4 Design Codes for Pressure Vessels

In modern competitive industry, new process plants are being set up rapidly and existing units are continually revamped, modernized and refurbished for the chemical, petrochemical, pharmaceutical, energy, refinery, and nuclear industries all over the globe. A variety of equipment is needed for the storage, handling and processing of hydrocarbons and chemicals in these processing plants. Unfired pressure vessels such as storage tanks, Horton spheres, mounded bullets, reactors, autoclaves, distillation/fractionating columns, and heat exchangers are some of the basic components of any such process plant.

Various codes specifying the requirements for the design, material, fabrication, inspection, and testing of pressure vessels have been written and adopted as national standards in various countries. Most the international pressure vessels have been developed with a higher degree of technical similarity between them. Core area such as vessel class, design criteria and requirements for independent inspection and certification are based on similar (but not identical) guiding principles.

A compilation of websites of various organizations, associations, technical standards, and current European Union (EU) legislation relating to pressure equipment sector is presented in [11.4].

The subsequent subsections present a brief discussion of the key features of some of the international pressure vessel codes.

11.4.1 ASME Boiler and Pressure Vessel Code

The ASME boiler and pressure vessel code, section VIII, published by the American Society of Mechanical Engineers, also known as ASME International, is a widely accepted code in the USA and 80 other countries in the world. The code is written by voluntary engineering talent and is constantly upgraded by the corresponding committee members to include the latest developments in material and design methodology.

ASME section VIII is written against a well-defined theoretical background and is divided into three subdivisions: 1, 2 and 3. The important design rules of both the codes are summarized in the respective appendices of each code. The contents of the three divisions are organized to cover specific pressure ranges, as illustrated in Table 11.4.

As can be seen from the code comparison table, division 2 permits higher working stresses at the expense of stringent material testing and more-careful quality control. In addition to detailed design rules division 2 provides for discontinuities, fatigue, and other stress analysis considerations, which are based on maximum shear stress theory. Division 2 also contains rules for vessels with layered construction (multiwalled vessels)

Fatigue Analysis

Clause AD-160 of division 2 provides several methods for exempting fatigue evaluations. If the cyclic operation does not meet all the conditions of AD-160 a fatigue evaluation method as per appendix 5 or 6 is added. The stress ranges are first determined for the specified cyclic operation and then, using fatigue curves presented in appendix 5, the associated number of allowable cycles and Miner's rule are used to determine the life fraction and cumulative damage. Fatigue curves up to 370 °C for carbon/low-alloy steel and 430 °C for austenitic stainless steel are provided in the code.

Division 3, which is a comparatively recent publication, provides a state-of-the art code and is intended for high-pressure applications where fatigue and fracture dominate.

11.4.2 PED Directive and Harmonized Standard EN 13445

The early phase of development of design codes and associated legislation in the pressure equipment sector was done predominantly at the national level. In view of the substantial economic potential of this sector, a need was felt in the European community to introduce a uniform and harmonized regulatory framework within the EU. The objective of the common legislation mechanism promulgated through directives was to minimize, if not eliminate, trade barriers between EU member states for pressure equipment, at the same time meeting essential safety requirements stipulated by the new directives.

Of the several European directives enforced since 1987, the Pressure Equipment Directive (PED) (97/23/EC) and the Simple Pressure Vessel (SPV) Directive (87/404/EEC) are the two principal instruments for the pressure equipment sector. The approach of the directives includes the identification of the product, prescription of the essential safety requirements (ESRs) to be met by technical standards, demonstration of conformity and CE marking. Technical standards adopted by the European Committee for Standardiza-

Table 11.4 Comparison of various divisions of ASME codes vis-à-vis EN13445

	Section VIII division I	Section VIII division II	Section VIII division III	EN 13445 part 3
First publication	< 1940	1968	1997	2002
Units adopted	řF and ksi			°C and N/mm^2
Pressure limits	Upto 3000 psig	No limits, usually + 600 psig	No limits, usually from 10 000 psig	For gas-ps up to 3000 bar and/or ps* V up to 3000 bar L For liquid-ps up to 1000 bar and/or ps* V up to 10 000 bar L
Design factor				
Tensile strength	3.5	3.0		Ferritic/normalized steel – 2.4 Austenitic steel – 3
Yield strength	1.5	1.5	Yield based with reduction factor for yield to tensile ratio less than 0.7	Ferritic/normalized steel – 1.5 Austenitic steel – 1.5 or 1.2 (depending on rupture elongation)
Average stress for 1% creep in 100 000 h	1.0	1.0		Guidelines for design in creep range are under development
Allowable stress calculated by	Committee and provided in Tables in section II C for individual grades and products			Designer
Testing groups	NDE requirement dependent on weld efficiency factor	Stringent NDE requirement	NDE requirement is even more stringent than Division 2	Classified from I to IV in decreasing extent of NDT
Hydrostatic test	1.3 × design pressure	1.25 × design pressure	1.25 × design pressure	Greater of 1.43 × allowable pressure and 1.25 × temperature-adjusted pressure

Table 11.5 EN 13445:2002 *Unfired pressure vessels* – a quick reference

Reference	Title	Contents/special features
EN 13445-1:2002	General	Scope, extent of testing with respect to weld joint coefficient, material grouping, etc.
EN 13445-2:2002	Materials	Materials listed include steels with sufficient ductility, cast iron and aluminium. Permitted or tabulated design stress are not provided. List of normative annexures are provided which include technical, inspection and delivery condition. Material listed in CE harmonized product standard can be used for that product. Material listed in CE harmonized material standard, if appropriate to the product. European approval of material (EAM) required for materials not listed in the harmonized standards. Particular material appraisal (PMA) required for material not listed in the harmonized standards or approved via EAM.
EN 13445-3:2002	Design	See Table 11.6 listing features for individual components.
EN 13445-4:2002	Fabrication	Provides weld designs, tolerances, production testing, post-weld heat treatment and repair requirements.
EN 13445-5:2002	Inspection and testing	Specifies nondestructive testing, pressure testing, marking and documentation requirement for noncyclic operation and special provisions for cyclic operations.
EN 13445-6:2002	Requirement for design and fabrication of pressure vessels and pressure parts constructed of spheroidal graphite cast iron.	
CR 13445-7:2002	Supporting standard for guidance on the use of the conformity procedures.	
prEN 13445-8	Additional requirement for pressure vessels of aluminium and aluminium alloys.	

Table 11.6 A quick summary of EN 13445 part 3

Type of analysis	Component	Special features
Design by formulae for non-cyclic loading (used in clauses 7 to 16) (full pressure cycles less than 500 cycles)	Dished ends	Additional formulae for knuckle and knuckle buckling
	Cones and conical ends	Based on limit analysis
	Opening in shells	Pressure area method
	Opening in flat ends	Replacement of section modulus
	Flanges	Modified Taylor forge method, alternate method also provided based on limit load analysis carrying out detailed assessment of flange–bolt–gasket system, useful for joints where bolt loads are monitored
	Weld joint efficiency	Linked to testing group I/II/III/IV
	Tubesheets	In addition to traditional method, new method using limit analysis approach in which edge loads and restraints are addressed, is provided Allowable tube loads are calculated
	Bellows	Covers both thin- and thick-walled bellows Rules are based on Expansion Joint Manufacturer's Association (EJMA) standards
	Rectangular vessels	Covers both unreinforced and reinforced vessels where outside stiffeners are attached
	Non-pressure local loads	Local loads are assessed by comparison with allowable loads calculated based on limit load analysis
Design by formulae for cyclic loading (used in clauses 17 and 18) (full pressure cycles exceeding 500 cycles)	Simplified fatigue analysis (clause 17)	Fatigue strength calculated using fatigue curves, which are more conservative than those used in detailed fatigue analysis for the unwelded region Table is provided for selecting stress factor for various design configurations
	Detailed fatigue analysis (clause 18)	Correction factors depending upon temperature, thickness, mean stress and surface finish are applied on fatigue curves Cumulative damage calculated as per Miner's rule Guidance provided on bending stress calculations due to misalignment of weld and recommendation given for weld-toe grinding
Design by analysis (covered by annexes B and C)	Direct route (annex B)	New route based on Eurocode (for steel structure), which overcomes the shortcomings of the familiar stress categorization method, addresses failure modes directly by way of design checks under the influence of *actions* (all imposed thermo-mechanical quantities including pressure, thermal and environmental) using partial safety factor depending on nature of single action or combination of actions
	Stress categorization route (annex C)	Based on categorization of stresses into primary, secondary and peak stress and comparing the same with specified limits, Stress classification table is provided for given component and its location depending upon the source load to help list and allocate stress category

tion (CEN/CENELEC), which provide means for the user to comply with the ESRs of the PED, are called *harmonized standards*.

EN 13445:2002 Unfired Pressure Vessel is a major harmonized product standard within the CEN pressure equipment portfolio. The standard utilizes expertise and best practices from across the European member states as well other internationally accepted standards. The adoption of the first issue of this standard, published in 2002, was preceded by discussions between experts who took nearly 10 years to achieve a major technical convergence.

The various sections of EN 13445 along with a brief description of the respective contents are tabulated in Table 11.5, while Table 11.6 summarizes the essential features of the design rules for various pressure vessel components, as covered in part 3 of EN 13445. An excellent presentation of the background to the rules of EN 13445 part 3 is available [11.5].

11.4.3 PD 5500

PD 5500:2000, the *specification for unfired fusion welded pressure vessels*, is a recent replacement for BS 5500:1997. It was issued under the status of a published document (PD) in anticipation of simultaneous release of the harmonized standard EN 13445 for pressure equipment in Europe.

Table 11.7 Summary of the salient features of ASME, PD 5500 and the AD code

Information	ASME section VIII division 1		PD 5500		AD Merkblätter	
	Part	Summary	Section	Summary	Section	Summary
Responsibilities	UG-99	Responsibilities listed for manufacturer and authorized inspector (AI)	1.4	Responsibilities for code compliance is on manufacturer	Druckbehälter VO	Authorized inspector to issue final vessel certification
Certification	UG-115 to 120	Manufacturer to have *certificate of authorization* to construct ASME stamped vessel	1.4	Code compliance is documented by use of form *X* issued by manufacturer and counter signed by AI	Druckbehälter VO	Authorized inspector to issue final vessel certification
Construction categories	–	–	3.4	Three categories with different material and NDT requirement are defined	HP 0	Four testing groups I, II, III and IV are defined
Joint types	UW-3	Defines A, B, C, D category with different NDT requirements; UW-12 gives joint efficiencies	5.6.4	Defines A and B welded joints, with different NDT requirements		
Weld joint details	UW-12, UW-13, UW-16	Shows typical weld joints for guidance and are not mandatory	E.1(1)–E.1(6)	Shows typical weld joints for guidance and are not mandatory		
Welder's approval	UW-28, 29, 48	Weld procedure specification (WPS), procedure qualification record (PQR) and welder's qualification are required	5.2, 5.3	WPS, PQR and welder's qualification are required	HP 2/1	WPS, PQR and welder's qualification are required
Permissible materials	UG-4, UG-10	UG-4 refers to materials specified in ASME section II	2.1.2	This sections references British standard materials	W0 to W13 for metallic and N series for non-metallic material	W0 to W13 cover all types of alloyed/unalloyed steel, castings, forgings, clad steel bolts and nuts but do not include gaskets
Material identification	UG-94	Material for all pressure parts is marked and certified for traceability	4.1.2	Positive material identification is required for all pressure parts		
NDT techniques	UW-51 to 53, UW-11, UW-42	NDT techniques are detailed in ASME section V	5.6.4	Ultrasonic testing (UT) and radiographic testing (RT) testing both are acceptable	HP 5/3	DIN standards are referenced for all NDT techniques
Assembly tolerance	UG-80, UW-33	Tolerances for circularity and alignment are specified	4.2.3	Tolerances for circularity and alignment are specified		
Pressure testing	UG-99, UW-50	Test pressures are specified along with other requirements of testing	5.8.1	Test pressures are specified along with other requirements of testing	HP 8/2, HP 30	
Heat treatment	UCS-56				HP 7/1 to HP 7/4	

Though PD 5500 does not currently have formal status as a harmonized standard and compliance with its technical requirement does not qualify presumption of conformity to PED, it does include information showing how its technical content would comply with the ESR of the directive.

Fatigue Analysis

Annexure C of the code presents criteria for exemption from fatigue analysis. A simplified fatigue analysis using design curves can be done using conservative estimates of the stress range due to pressure changes and thermal gradients. By using appropriate an design curve

to obtain allowable cycles and satisfying a cumulative damage rule, a simplified analysis can be done.

A detailed analysis is required if the specified criterion is not satisfied. Detailed methods for determining stresses due to pressure, thermal gradient, and piping loads based on using stress concentration factors are provided in annexure G. Based on the individual stresses thus obtained, the maximum principal stress range for each individual cycle is determined and fatigue evaluation is done as per annexure C.

The fatigue curves are limited to 350 °C for ferritic steel and 430 °C for austenitic stainless steels.

11.4.4 AD Merkblätter

Arbeitgemeinschaft Druckbehälter (AD) Merkblätter regulations are a set of generally accepted rules of technology regarding pressure vessels and contain safety requirement for equipment, design, manufacture and testing, and materials. These regulations are compiled by seven trade associations of Germany, who together form the AD. The AD associations represent a balanced combination of material and pressure vessel manufacturer's, operators, employer's liability insurance, and technical inspectorates.

To a large extent the AD Merkblätter code is based on Deutsches Institut für Normung (DIN) standards and is continuously amended in keeping with technical progress. The associations has published a regulation called AD 2000 conforming to the safety requirement and other stipulations as laid down in the PED, which has been made compulsory in Europe from May 2002 onwards.

A quick summary showing the key technical points of PD 5500 in comparison with the ASME section and AD Merkblätter code are shown in Table 11.7.

11.5 Heat Exchangers

Heat exchangers are devices that transfer heat from a hot to a cold fluid. They are used extensively in processing plants and are given specific names when they serve a special purpose, for example superheaters, condensers, evaporators etc. In a surface-type exchanger, the two process fluids are separated by a physical barrier and the heat is transferred from the warm/hot fluid through the barrier to the cold fluid. Shell-and-tube and plate exchangers are the most commonly used types of surface exchangers.

The construction of shell-and-tube exchangers is broadly divided into the shell side and tube side (also called the channel section). Several small-diameter tubes (on the tube side) are attached to larger pressure vessels (known as the shell side), or parts thereof, called tube sheets. The tubes are distributed within the tube

Table 11.8 Design rules for tube sheets as per different codes

	TEMA	ASME section VIII division 1	EN 13445-3
U tube tubesheet			
Reference section	Section 1999 edition	Appendix AA-1 edition 2001 and UHX-12 edition 2002 (mandatory)	Clause 13 D-2002 (based on Code francais de construction des Appareils a Pression (CODAP)/unifired pressure vessel (UPV) rules)
Assumptions	Perforated tubesheet and unperforated rim not accounted for; the effect of tube sheet attachment with shell/channel not considered	Refined and rational analytical treatment is used after taking into account actual geometry (based on model proposed by F. Osweiller in 2000)	
Ligament efficiency	$0.45 \leq \eta \leq 0.60$	$0.25 \leq \mu \leq 0.35$	
Safety factor	2.6	1.5	2.0
Allowable stress		Stress classification of division 2 appendix 4	Based on primary and secondary stress as per appendix C
Remarks	TEMA approach leads to lower thickness than ASME due to higher ligament efficiency and high safety factor	Higher thickness obtained in ASME rules than UPV/CODAP due to lower allowable stress	

Table 11.8 (cont.)

Fixed/floating tubesheet			
Reference section		Appendix AA-2 edition 1992	Clause 13 E-2002 (based on CODAP/UPV rules)
Assumption	Stiffening effect of tube bundle and weakening effect of tube holes are assumed to counterbalance each other, coefficient F is not dependent on the stiffness ratio X of the axial tube bundle rigidity to the bending rigidity of the tube sheet	Coefficient F is dependent on the stiffness ratio X	Coefficient F is dependent on the stiffness ratio X; the value of F is higher than ASME since tubes are assumed to be uniformly distributed over the whole tube sheet
Remarks	TEMA does not provide the same design margin for all cases, leading to over-thickness for higher X and under-thickness for lower X		

sheet in a certain pattern, the three most common of which are the triangular, square and rotated triangular, and square.

Designs for the joint between the tube and tube sheet vary widely and are chosen to be compatible with the severity of the service conditions. The joint may be expanded, welded, or a combination of both. There are various constructional details of welded joints, the choice of which is based upon service and environment.

Various configurations and designs of shell-and-tube exchangers are covered extensively by the Tubular Exchanger Manufacturers' Association (TEMA).

The principle of mechanical design for most of the components of the heat exchanger is identical to the design of a pressure vessel. However, the design of the tube sheet is typically different because of its constructional geometry. The tube sheet design rules have been rationalized in recent years by various pressure vessel codes, which are summarized in the Table 11.8.

11.6 Material of Construction

11.6.1 Carbon Steel

Steel is an iron alloyed with carbon at a level of 0.05–2.0%. In addition it contains smaller proportions of phosphorus, sulphur, silicon, aluminum, and manganese. These steels are known as plain carbon steels, which are classified as mild-, medium- and high-carbon steels according to the percentage of carbon.

Mild steel has 0.05–0.3% carbon by weight, medium-carbon steel has a carbon content of 0.3–0.6% and high-carbon steel has more than 0.5%, up to a maximum of 2%, carbon content. Mild steels are the most versatile materials for the construction of pressure vessel due to their good ductility and relative ease of forming, rolling, forging, fabrication, and welding. They are also most suited and economic for applications where the rate of corrosion is low. During the manufacture of these steels, silicon and/or aluminum are added to react with dissolved oxygen in the molten metal alloy to form a slag of Al_2O_3/SiO_2, which floats to the top and is removed; the resultant steel is called *killed steel*. A fully killed steel usually contains less than 150 ppm oxygen and at least 0.10% silicon. Besides being cleaner due to the formation of fewer oxides and inclusions, fully killed steels are much easier to weld due to a lower tendency to outgas during welding.

Welding

Mild-steel electrodes are grouped into those with rutile-type flux covering and those with low hydrogen flux covering. Rutile-covered electrodes are used for general fabrication involving thinner sections, lower tensile strength, and in applications where there is no requirement for impact properties. For all other applications, where strength, impact properties, and weld quality are essential, low-hydrogen-type electrodes are employed. The flux covering eliminates sources of potential hydrogen and thus minimizes the chances of clod cracking.

11.6.2 Low-Alloy Steel

Low-alloy steel contains additions of the elements Ni, Cr, Si, Mo, and Mn in amounts totaling less than 5%. The added elements improve mechanical properties, heat treatment response and /or corrosion resistance. The weld ability of low-alloy steels is also good; however since these steels are hardenable by heat treatment, they do require careful attention to welding procedure including pre- and post-weld heat treatment (PWHT) for stress relief, which is discussed later.

The workhorse alloys for pressure vessels, exchanger and heater tubes, and piping for elevated temperature service, usually greater than 250 °C, contain 0.5%–9.0% chromium plus molybdenum. With the increasing chromium content, resistance to high-temperature hydrogen attack, and resistance to sulfidation and oxidation increases.

Developments

There has been an increasing trend towards improved toughness properties and temper embrittlement resistance by restricting levels of impurity elements at the ladle.

To minimize the temper embrittlement of low-alloy steel, the phosphorous content is restricted to 0.010% or lower, while the combined phosphorous–tin content is limited to 0.010%.

The effects of the tramp elements have been addressed by the Bruscato factors X and J, as defined by the following equations:

$$X = \frac{(10\mathrm{P} + 4\mathrm{Sn} + 5\mathrm{Sb} + \mathrm{As})}{100}$$

(all elements in ppm) ,

$$J = (\mathrm{Si} + \mathrm{Mn})(\mathrm{P} + \mathrm{Sn}) \times 10^4 \quad \text{(elements in \%)} .$$

For this reason PWHT is a must for all creep-resisting Cr–Mo-type alloy steels. PWHT also stabilizes and softens the microstructure at the heat-affected zone (HAZ) of the weld. If the base metal is quenched and tempered, higher PWHT temperatures can be specified for improved resistance to creep embrittlement.

Fabrication

In view of the criticality of preheating, post-heating, and PWHT, it is prudent to employ specialized heat-treatment techniques using electrical resistance pads or induction coils,especially for thick piping sections.

11.6.3 NACE standards

The National Association of Corrosion Engineers (NACE) is a worldwide technical organization that studies various aspects of corrosion in refineries, chemical plants and other industrial systems.

NACE standard MR0175, entitled *Sulfide corrosion cracking resistant metallic material for oilfield equipment*, is widely used for applications in sour gas and oil environments. NACE compliance is recommended in systems where there is a likelihood of sulfide cracking due to the presence of a measurable amount of H_2S. Since the susceptibility of carbon, low-alloy and austenitic stainless steel to sulfide corrosion is directly related to strength and hardness level, it these standards recommend that the hardness of the aforementioned plate material should be restricted to 22 HRC (200 BHN). The cold working of these steels during forming/rolling shall be less than 5%. Post-weld heat treatment is to be carried out for carbon steels in the case of greater cold working. A few duplex stainless and some nickel-based alloys are also acceptable according to the NACE criteria, subject to a maximum hardness level of 28 and 35 HRC, respectively.

Welding

All weld procedures must be qualified to meet the same hardness levels standards as specified for the corresponding parent material.

11.6.4 Comparative Standards for Steel

The discussion for carbon steel material and the discussion to follow for other construction materials are generally based on the generic composition of materials without referring to any code of construction. It is difficult to furnish equivalence of any grade of material from one code to another; at best steels grades can be compared based on the closest matching technical requirements. Tables 11.9, 10 list some of the comparable standards for flat and tubular products commonly used in fabrication of pressure vessel.

11.6.5 Stainless Steel

Alloys of iron and carbon with over 12% chromium, which resist rusting under most atmospheric conditions, are called stainless steels. The alloys become more corrosion resistant as the chromium level increases. As more nickel is added, several phases are possible and the alloy may undergo transformation depending on the

Table 11.9 Comparative chart for material standards of flat products (plates) for pressurized use

Harmonized European standard			ASTM	DIN	BS 1501
Description	**Standard**	**Grade**			
General requirement	EN 10028-1		SA 20		
				DIN 17155	
Non-alloy and alloy steel with elevated-temperature properties	EN 10028-2	P235GH	SA 283 Gr C/ SA 516 Gr 55	H I	161 Gr 360/ 164 Gr 360
		P265GH	SA 516 Gr 60	H II	161 Gr 400/ 164 Gr 400
		P295GH	SA 516 Gr 65	17Mn4	224 Gr 490
		P355GH	SA 414 Gr G	19Mn6	
		16Mo3	SA 204 Gr B	15Mo3	1503-243 B
		13CrMo4-5	SA 387 Gr 12	13CrMo4 4	620 Gr 27
		10CrMo9-10	SA 387 Gr 22	10CrMo9 10	620 Gr 31
				DIN 17102	
Weldable fine-grain, normalized	EN 10028-3	P275N		StE 285	224 Gr 400
		P275NH	SA 516 Gr 60	WStE 285	224 Gr 430
		P275NL1	SA 662 Gr A	TStE 285	
		P275NL2		EStE 285	
		P355N	SA 537 CL 1	StE 355	224 Gr 490
		P355NH	SA 662 Gr C	WStE 355	224 Gr 490
		P355NL1	SA 737 Gr B	TStE 355	224 Gr 490
		P355NL2		EStE 355	
		P460N		StE 460	
		P460NH		WStE 460	
		P460NL1		TStE 460	
		P460NL2		EStE 460	
				DIN 17280	
Nickel alloy steel with low-temperature properties	EN 10028-4	11MnNi5-3		13MnNi6	
		12Ni14	SA 203 Gr D,E,F	10Ni14	
		X12Ni5 (12Ni19)	SA 645	12Ni19	
		X8Ni9 NT	SA 353		
		X8Ni9 QT	SA 553		
Weldable fine-grain, thermomechanically rolled	EN 10028-5	P355M / ML1/ML2			
		P420M / ML1/ML2			
		P460M / ML1/ML2			
Weldable fine grain, quenched and tempered	EN 10028-6	P690Q / QH QL1/QL2	SA 517, SA 533, SA724		
Stainless steel	EN 10028-7	Various gardes (refer to Table 11.10)	SA 240	DIN 17440 DIN 17441	BS 1449-2 and BS 1501-3

Note: Common legend for Tables 11.9, 10 is shown under Table 11.10

Table 11.10 Comparative chart for material standards of tubular products for pressurized use

Harmonized European standard			ASTM		DIN	BS		
Description	**Standard**	**Grade**						
Non-alloy for general use	EN 10216-1	P195	SA 53 Gr A/SA 106 Gr A			BS 3059-1 Gr 320		
		P235	SA 53 Gr B/SA 106 Gr B		DIN 1629 St 37	BS 3601 Gr 360		
		P265			DIN 1629 St 44	BS 3601 Gr 430		
					DIN 17175	**BS 3059-2 Carbon/alloy**	**BS 3602 C–Mn**	**BS 3606 Exchanger tube**
Unalloyed and alloyed for elevated temperature	EN 10216-2	P195GH	SA 179 – Cold-drawn tubes for exchanger					Gr 320
			SA 192 – Boiler tubes					
		P235GH			St 35.8	360	360	400
		P265GH	SA 210 Gr A1 – Boiler/superheater tube		St 45.8	440	430	440
			SA 210 Gr C		17Mn4			
			Tube	**Pipe**				
		16Mo3	SA 209 T1	SA 335 P1	15Mo3	243		243
			SA 213 T2	SA 335 P2			BS 3604 (ferritic)	
		10CrMo5-5	SA 213 T11	SA 335 P11			621	621
		13CrMo4-5	SA 213 T12	SA 335 P12	13CrMo44		620	620
			SA 213 T21	SA 335 P21				
		10CrMo9-10	SA 213 T22	SA 335 P22	10CrMo9 10		622	622
		X11CrMo5	SA 213 T5	SA 335 P5			625	625
		X11CrMo9-1	SA 213 T9	SA 335 P9	12CrMo195	620-470	620-470	
		X10CrMoVNb9-1	SA 213 T91	SA 335 P91	X10CrMoVNb91	91		
					DIN 17179			
Unalloyed and alloyed, fine grain	EN 10216-3	P275NL1/NL2			WStE/TStE 285			
		P355N/NH			EStE/StE 355			
		P355NL1/NL2			WStE/TStE 355			
		P460N/NH			EStE/StE 460			
		P460NL1/NL2			WStE/TStE 460			
			ASTM 333	**ASTM 334**	**DIN 17173**	**BS 3603 carbon/alloy**		
Unalloyed and alloyed, low temperature	EN 10216-4	P215 NL	Gr 1	Gr 1	TTSt35 N			
		P265 NL	Gr 6	Gr 6		430 LT		
		12Ni14	Gr 3	Gr 3	10Ni14	530 LT		
		12Ni14+ QT	Gr 3	Gr 3	10Ni14	530 LT		
		X10Ni9	Gr 8	Gr 8	X8Ni9	509 LT		
		X10Ni9 + QT	Gr 8	Gr 8	X8Ni9	509 LT		

Legend for steel grade to EN series
G – Other characteristics follows, N – Normalized condition, H – Elevated temperature property required, M – Thermo mechanically rolled, QT – Quench and tempered, L1 – Low temperature property, impact testing at −50 °C, L2 – Special low temperature property, impact testing at −50 °C, enhanced requirement

actual content of Cr–Ni–Fe and C. Gamma or austenitic stainless steel is an iron alloy containing at least 18% Cr and 8% Ni with carbon up to 0.10%. The austenitic structure provides a combination of excellent corrosion resistance, oxidation, and sulfidation resistance with high creep resistance, toughness and strength at temperatures up to 550 °C. The 18Cr–8Ni stainless alloys form a series known as the American Iron and Steel Institute (AISI) type 300 series, with varying amounts of carbon, molybdenum, and titanium added.

The 18–8 series have good formability besides being readily weldable without stress relief; however they can be hardened by cold working. These steels are susceptible to grain-boundary chromium carbide precipitation, known as *sensitization*, when subjected to a temperature range of 535–800 °C. To prevent sensitization, low-carbon grades (C < 0.03%) and stabilized grades with added columbium or titanium are used.

The traditional grades of stainless steel that are extensively used in the fabrication of chemical plant and equipment are AISI 304 type and the 2% molybdenum-bearing 316 types along with their low-carbon versions 304L and 316L or the stabilized grades 321 (with Ti) or 347 (with Cb) where intercrystalline corrosion is to be avoided.

Higher-alloyed grades such as AISI 309 and 310 have a higher chromium content that makes them suitable for high-temperature applications such as furnace liners, preheaters and column trays.

Developments

Newer grades of stainless steel have been developed to overcome the limitations of low 0.2% proof stress, sensitivity to stress and pitting corrosion especially in chloride media, inadequate corrosion in reducing media, and preferential attack on the ferrite phase in strong oxidizing media.

Nitrogen alloyed steel such as AISI 304LN and 316LN have been developed with the addition of 0.2% nitrogen, resulting in improved proof stress by about 15%. Nitrogen alloyed steel finds wide application in the transportation and storage industries.

For reactors, strippers, and condensers in urea service steels with higher chromium and nickel contents with nil ferrite have been developed. Modified compositions of AISI 310 steel such as SANDVIK 2RE69 and Assab 725 LN have been developed for strong oxidizing conditions in fertilizer plants.

For elevated-temperature applications in furnaces, and hydrocarbon and steam reformers, where higher creep strength is also necessary, casting alloys such as HK-40 and IN657 have been developed.

Duplex stainless steels were developed to combine the attractive properties of ferritic and austenitic stainless steels. In simple terms, the ferrite could be said to provide mechanical strength and stress corrosion resistance, while the austenite provides ductility and the two combine to produce a fine-grained, two-phase microstructure with high strength and good corrosion resistance. Of the many alloying elements Cr and Mo enhance the formation of ferrite, while Ni and N stabilize the austenite. Resistance to pitting and crevice corrosion in chloride environments is increased, expressed by the pitting resistance equivalent $PRE_N = \%Cr + 3.3 \times \%Mo + 16 \times \%N$.

This number is used to rank materials according to their expected resistance to pitting corrosion. A 23% Cr Mo-free grade would have a PRE_N value of about 25. Regular Mo-alloyed duplex grades have a PRE_N value of 30–36. Steels having a PRE_N value higher than 40 are normally defined as super-duplex stainless steels.

Welding

Duplex stainless-steel welds with matching composition filler material show high ferrite levels, which has low toughness and poor ductility. Therefore all welding consumables for duplex materials are over-alloyed with nickel, which allows more austenite to form so that the ferrite level in welds is lowered and the welds have good ductility and corrosion resistance. It is recommended to achieve a ferrite content of approximately 22–70%; equivalent to 30–100 FN (ferrite number). In addition to the ferrite count, a corrosion test in ferric chloride is also carried out as per ASTM G48, which gives a good assessment of the corrosion resistance of the weld metal. As per this test, the critical pitting temperature (CPT) is specified at 22 °C for duplex and 35 °C for super-duplex steel.

Preheating of duplex material is not required except where heavy loads on high-ferrite-containing welds may cause cracking. Post-welding solution annealing is required only in cases where the resultant weldment has deteriorated due to detrimental phase transformation and/or has high ferrite levels.

Root passes of nitrogen-alloyed stainless steel are welded with higher-alloyed filler, to compensate for the influence of nitrogen.

Urea-grade stainless steels and steels in high-temperature applications are welded with matching composition electrodes enriched with 4–5% manganese to counteract the tendency for microfissuring.

11.6.6 Ferritic and Martensitic Steels

Ferritic steels are chromium–iron stainless steel with little or no nickel and form a body centric structure unlike the face-centered austenitic steel. When ferritic steels are modified by heat treatment, they become hardened and form *martensitic* steels.

Martensitic steels derive their excellent hardness from the high levels of carbon added to their alloy. The most commonly used martensitic steel is ASTM type 410 stainless steel used for column tray and tower lining in crude service for refinery applications. The increased carbon level in 410 steel results in a much harder martensitic cutlery steel or tool steel type 420.

By increasing the percentage of chromium, transformation hardening is suppressed, as in ASTM types 430 and 446, which are essentially ferritic. These steels are resistant to chloride stress corrosion cracking; however they are subject to ductile–brittle temperature embrittlement, thereby restricting their mechanical properties.

The limiting values for X and J factor are usually specified for welding consumables, although it would be preferable also to restrict these for the base material.

A step-cooling simulation treatment is performed on higher-thickness quench and tempered plates to determine susceptibility to embrittlement phenomenon in terms of meeting the specified shift in the 40 ft – lb charpy V-notch (CVN) transition temperature.

Welding

With increasing base and weld material strength and hardenability, hydrogen diffusibility in the weldment is kept below 5 ml/100 g of deposited weld metal. The flux covering of all electrodes is of the low-hydrogen type and employs binders that give high resistance to moisture absorption. Prior to use, they are dried or baked at the manufacturer's recommended temperature.

Although there are several theories for determining the optimum preheat temperature, the common industrial practice is to use carbon equivalent as guidance to select the temperature, as shown here

CE	Temperature °C
< 0.4	50
< 0.55	100
< 0.70	150
< 0.8	200
< 0.9	250

Maintaining the preheat after welding (also called post-heating) for a specified period (generally 300–350 °C for 30 min) in some cases (say for pipe thicknesses greater than 12 mm with a chromium content of 2–7%) also helps to reduce hydrogen levels, thereby preventing cold cracks and stress corrosion cracking.

Martensitic steels can be welded but caution needs to be exercised as they will produce a very hard and brittle zone adjacent to the weld. Cracking in this zone can occur (particularly in thicker sections) and therefore preheating and PWHT is recommended.

Though ferritic steels are less prone to cracking due to their lower strength and non-hardenability, the weldment suffers from excessive grain growth, sensitization and a lack of ductility. Due to the excessive grain growth problem, only thin-gauge sheets are generally used. Filler material can be of either a similar composition, or alternatively an austenitic grade can be used to help weld toughness and increase ductility.

Table 11.11 lists some of the comparable standards for stainless-steel grades commonly used in the fabrication of pressure vessel.

Developments

Since the ferritic grades do not possess good welding properties, hybrid utility ferritics such as 3CR12 with controlled martensite (dual phase steel) have been developed to overcome these welding difficulties. Using new steel-refining techniques, along with the addition of titanium or niobium, it has been possible to develop extremely corrosion-resistant grades such as *superferritic stainless steel*.

11.6.7 Copper and Nickel Base Alloys

Brass

Brasses are commercially produced with varying percentages of copper and zinc to provide a range of properties depending on the end-use requirements.

Admiralty brass, which is widely used for tubes in water-cooled condensers for low water speeds, is an alloy brass containing 71% Cu, 28% Zn, and 1% Sn. To prevent dezincification, small amounts of arsenic, phosphorous or antimony are added.

At high water speeds and when seawater contains air bubbles, aluminum brass containing 2% Al is more suitable due to the formation of a protective film. For tube plates in condensers and exchangers, it is usual practice to use high-zinc brasses, hot-rolled Muntz metal (60% Cu, 40% Zn) or Naval brass (60% Cu, 39% Zn,

Table 11.11 Comparative chart for various stainless-steel grades

Structure	Hardenability	ASTM 240 Grade	UNS No.	EN 10028-7 Grade	Number	Analysis built up from basic type
Austenitic	Hardenable by cold work	304	S 30400	X5CrNi18-10	1.4301	Cr 18% + Ni 8% basic type
		304L	S 30403	X2CrNi18-10	1.4303	304 with low carbon
		308	S 30800	X2CrNiMo8-14-3	1.4432	Higher Cr and Ni for more corrosion and scaling resistance
		309	S 30900	X2CrNiCuWN25-7-4	1.4501	Still higher Cr and Ni
		310	S 31000	X5CrNi25-21	1.4335	Highest Cr and Ni (Cr 25% + Ni 20%)
		316	S 31600	X5CrNiMo17-11-2	1.4401	Mo added for corrosion resistance
		316L	S 31603	X5CrNiMo17-12-2	1.4404	316 with low carbon
		316N	S 31603	X2CrNiMoN17-11-2	1.4406	316 with nitrogen added for low-temp. service
		321	S 32100	X6CrNiTi18-10	1.4541	Ti added to avoid carbide precipitation
		347	S 34700	X6CrNiNb18-10	1.4550	Cb added to avoid carbide precipitation
Martenstitic	Hardenable by heat treatment	410	S 41000	X12Cr13	1.4006	Cr 12% basic type
		420	S 42000	X46Cr13	1.4034	Higher C, cutlery application
		431	S 43100	X4CrNiMo16-5-1	1.4418	Higher Cr and Ni added for improved ductility
Ferritic	Non-hardenable	405	S 40500	X6CrAl13	1.4402	Al added to Cr 12% to prevent hardening
		430	S 43000	X6Cr17	1.4016	Cr 17% basic type
		442	S 44200			Higher Cr to resist oxidation and sulfidation at higher temperature
		446	S 44600			
Precipitation-hardening steel	Age-hardenable	17-7 PH	S 17700	X7CrNiAl17-7	1.4568	
		14-8MoPH	S 13800	X8CrNiMoAl15-7-2	1.4532	

and 1% Sn) to take advantage of higher tensile strength, although the two-phase structure of these alloys cannot be satisfactorily inhibited against dezincification.

Bronze

Bronze is a tin alloy of copper with other elements such as aluminum added for additional properties. Because of the hardening effect of tin, hot-rolled bronze plates have greater strength than brass plates and therefore can be used for tube plates and channel material for exchangers.

Cu–Ni Alloys

Alloys of copper and nickel have historically been used in saltwater condensers as they show better resistance to saltwater than brasses. Increasing nickel content was found to be beneficial and 30% Ni alloy was adopted as the standard for naval vessels. The addition of iron and Mn was found to improve resistance to impingement attack.

These alloys are used as tubes for heat exchangers, saltwater pipelines, and hydraulic lines as well as for several applications in marine and offshore

platform services. Cupronickel tubes are superior to brass in terms of better resistance to dezincification for applications involving higher metal temperature of water-cooled exchangers. They are also excellent materials for tube plates and, because of their good formability and weldability, they can be used in sheet form for the fabrication of heat-exchanger shells and water boxes. Monel, a nickel-copper alloy (67% Ni, 30% Cu) has good resistance to saltwater, and to hydrochloric and hydrofluoric acid under nonoxidizing conditions. Therefore they are excellent material for cladding and trays in towers handling acid vapors.

Other Nickel-based alloys are classified as:

1. Chromium bearing as in Inconel 600, and Hastelloy C-22 and C-276;
2. Containing chromium and molybdenum such as Inconel 625, Hastelloy B, Incoly 825;
3. Precipitation hardening alloys such as Monel K-500, and Inconel 817.

These alloys show excellent resistance to pitting, stress corrosion cracking in chloride environments and retain strength at elevated temperatures. Therefore they are excellent candidates for exchanger tubes, heat-transfer plates for plate heat exchangers (PHE), and pressure coils for steam/hydrocarbon reformers, naphtha-cracking furnaces etc. in the petrochemical and fertilizer industries.

Welding

Gas tungsten arc (GTAW), gas metal arc (GMAW), and shielded metal arc (SMAW) processes are most commonly used for welding brasses and bronze. Whereas thin gauges are welded with GTAW using zinc-free fillers, heavier gauges are joined by a GMAW/SMAW process using zinc-free silicon bronze or aluminum bronze fillers/electrodes. Zinc-free fillers are prescribed since the evolution of zinc fumes makes the weld porous and affects visual observation of the welding process; moreover zinc fumes are extremely hazardous to health. Argon and helium, either individually or in combination, are used for shielding in the case of the GTAW/GMAW process.

Welding consumables for welding of all Cu–Ni-based alloys are available with compositions matching the specific parent material with the generous addition of manganese and/or niobium, which are added intentionally to give resistance to hot cracking and to raise hot strength. Most of these consumables are often used for dissimilar metal welding between the nickel base and most steels or between other ferrous and nonferrous alloys.

Carbon and silicon are controlled to low levels to minimize detrimental precipitates in the weld metal and HAZ for electrodes with specifications matching with some of the high-molybdenum alloys such as Hastelloy C276 and Hastelloy B.

References

11.1 J.F. Harvey: *Theory and Design of Pressure Vessels* (Van Nostrand Reinhold, Amsterdam 1985)

11.2 D.R. Moss: *Pressure Vessel Design Manual* (Gulf, Houston 1987)

11.3 H.H. Bednar: *Pressure Vessel Design Handbook* (Van Nostrand Reinhold, Amsterdam 1986)

11.4 C. Matthews: *Engineer's Guide to Pressure Equipment – The Pocket Reference* (Professional Engineering, Suffolk 2001)

11.5 G. Baylac, D. Koplewicz (Eds.): *EN 13445 Unfired Pressure Vessels – Background to the Rules in Part 3 Design* (UNM, Paris 2002), (Issue 2 download from www.unm.fr)

12. Turbomachinery

Meinhard T. Schobeiri

The following chapter consists of two sections. Section 12.1 presents a concise treatment of the theory of turbomchinery stages including the energy transfer in absolute and relative systems. Contrary to the traditional approach that treats turbine and compressor stages of axial, radial or mixed configurations differently, these components are treated from a unifying point of view.

Section 12.2 is dedicated to steady and unsteady performance of gas turbine engines, where the components are treated as generic modules. Thus, any arbitrary power generation or aircraft gas turbine engine with single or multiple shafts can be composed of these modules. Several examples show, how different gas turbine configurations can be constructed and dynamically simulated. Finally, a section about the new generation gas turbines shows, how the efficiency of gas turbines can be improved far beyond the existing level.

This chapter is based on [12.1], where the reader finds detailed explanation of relevant aerodynamic aspects of turbomachines, their component losses and efficiencies, and the design and off-design performance calculations.

12.1 Theory of Turbomachinery Stages

12.1.1 Energy Transfer in Turbomachinery Stages

Energy transfer in turbomachinery is established by means of a number of stages. A *turbomachinery stage* consists of a row of fixed, guide vanes called *stator blades*, and a row of rotating blades termed the *rotor*. To elevate the total pressure of a working fluid, *compressor stages* that partially convert the mechanical energy into potential energy are used. According to the law of conservation of energy, this energy increase requires

an external energy input which must be added to the system in the form of mechanical energy. Figure 12.1 shows a schematic of an axial compressor stage that consists of one stator and two rotor rows. In general, a *compressor component* starts with a rotor row followed by a stator row. Compressor configurations that start with an *inlet guide vane* are also found. To define a unified station nomenclature for the compressor and turbine stages, we identify station number 1 as the inlet of the stator, followed by station 2 as the rotor inlet, and station 3 as the rotor exit. The absolute and relative flow angles are counted counterclockwise from a horizontal line. This convention allows an easier calculation of the off-design behavior of compressor and turbine stages during a transient operation, as we will see later. Different angle conventions are used in the literature [12.2–5]. The working fluid enters the first rotor with an *absolute velocity* in the axial direction (Fig. 12.1b), where it is deflected in the direction of the rotor's leading edge. As a result of the rotational motion of the rotor, a major part of the mechanical energy input is converted into the potential energy of the working medium, causing the total pressure to rise. During the compression process, the absolute velocity within the stator and the relative velocity vector within the rotor decrease. To convert the total energy of a working medium into mechanical energy, *turbine stages* are used. Figure 12.2 exhibits a turbine stage within a multistage environment. As shown, the mean diameter may change from inlet to exit. The continuous increase in flow path cross section is due to the continuity requirement.

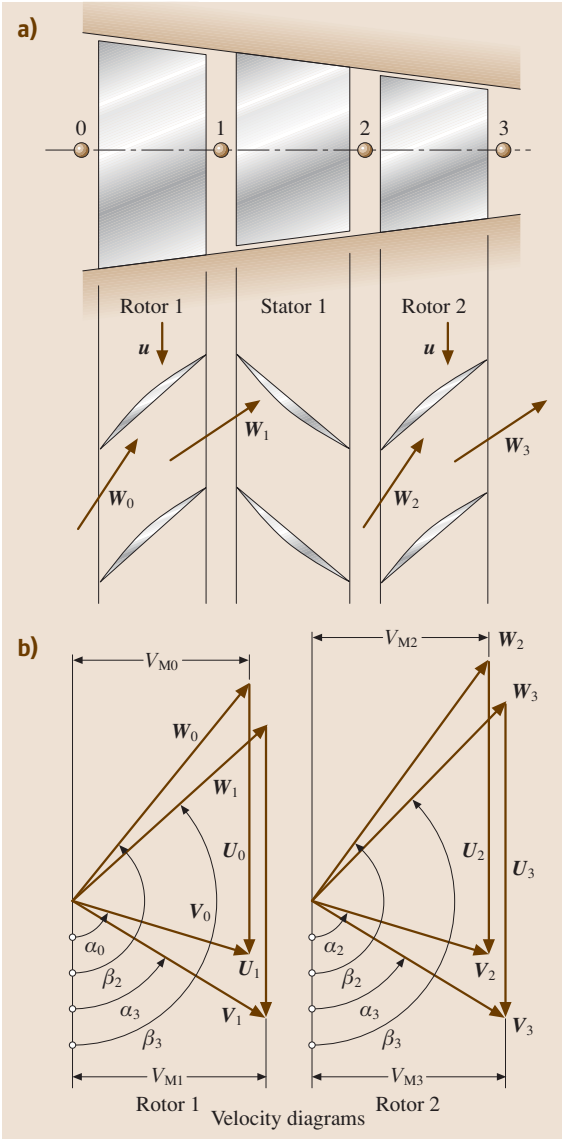


Fig. 12.1 (a) An axial compressor stage with a rotor–stator–rotor configuration and (b) velocity diagrams for the first and second rotor

12.1.2 Energy Transfer in Relative Systems

Since the rotor operates in a relative frame of reference (relative system), the energy conversion mechanism is quite different from that of a stator (absolute system). A fluid particle that moves with a relative velocity $\boldsymbol{W}$ within the relative system that rotates with the angular velocity $\boldsymbol{\omega}$, has an absolute velocity

$$\boldsymbol{V} = \boldsymbol{W} + \boldsymbol{\omega} \times \boldsymbol{R} = \boldsymbol{W} + \boldsymbol{U}, \quad \boldsymbol{\omega} \times \boldsymbol{R} = \boldsymbol{U}\,, \tag{12.1}$$

where $\boldsymbol{R}$ is the radius vector of the particle in the relative system. Introducing the absolute velocity vector $\boldsymbol{V}$ in the equation of motion [12.1, Chap. 3, Eq. (3.37)] and multiplying the results with a relative differential displacement $\mathrm{d}R$, we get the energy equation for an adiabatic steady flow within a relative system

$$\mathrm{d}\left(h + \frac{1}{2}W^2 - \frac{\omega^2 R^2}{2} + gz\right) = 0 \tag{12.2}$$

or the relative total enthalpy

$$H_\mathrm{r} = h + \frac{1}{2}W^2 - \frac{\omega^2 R^2}{2} + gz = \text{const}\,. \tag{12.3}$$

Neglecting the gravitational term, $gz \approx 0$, (12.3) can be written as

$$h_1 + \frac{1}{2}W_1^2 - \frac{1}{2}U_1^2 = h_2 + \frac{1}{2}W_2^2 - \frac{1}{2}U_2^2\,. \tag{12.4}$$

Equation (12.4) is the equation for the energy transformed in a relative system. As can be seen, the transformation of kinetic energy undergoes a change while the transformation of static enthalpy is frame indifferent. With these equations together with the energy balance [12.6] we can analyze the energy transfer within an arbitrary turbine or compressor stage.

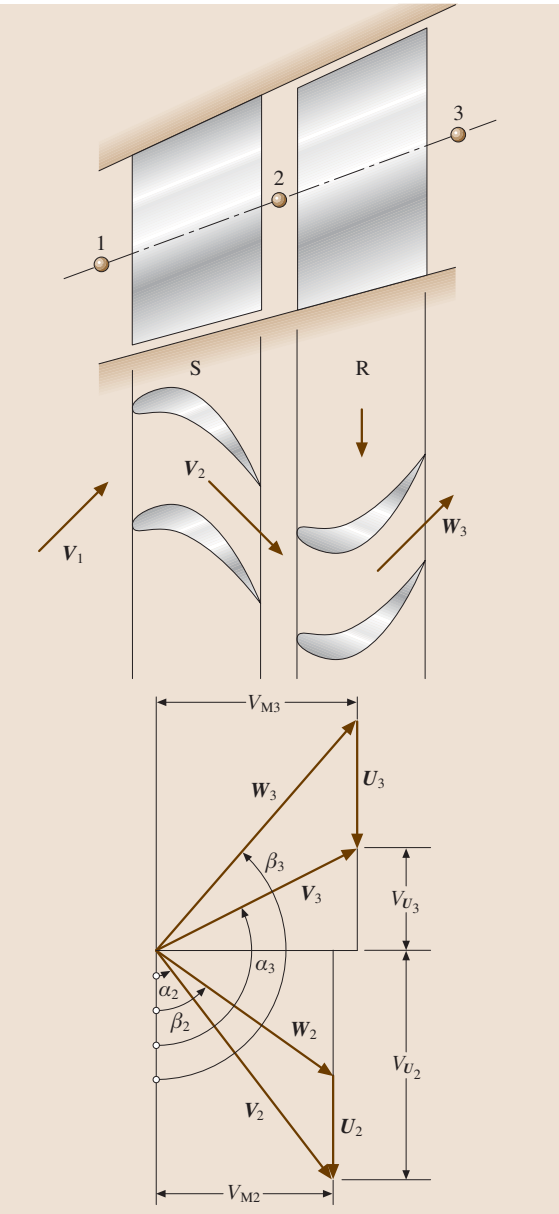


Fig. 12.2 An axial turbine stage with velocity diagram

12.1.3 General Treatment of Turbine and Compressor Stages

In this chapter, compressor and turbine stages are treated from a unified physical point of view. Figures 12.3 and 12.4 show the decomposition of a turbine and a compressor stage into their stator and rotor rows. The primes $'$ and $''$ refer to stator and rotor rows, respectively. As seen, the difference between the isentropic and the polytropic enthalpy difference is expressed in terms of dissipation $\Delta h'_{\mathrm{d}} = \Delta h'_{\mathrm{s}} - \Delta h'$ for turbines and $\Delta h'_{\mathrm{d}} = \Delta h' - \Delta h'_{\mathrm{s}}$ for compressors. For the stator, the energy balance requires that $H_2 = H_1$. This leads to

$$h_1 - h_2 = \Delta h' = \frac{1}{2}\left(V_2^2 - V_1^2\right). \tag{12.5}$$

Moving to the relative frame of reference, the relative total enthalpy $H_{\mathrm{r}2} = H_{\mathrm{r}3}$ remains constant. Thus, the energy equation for the rotor is according to (12.4) (Fig. 12.4)

$$h_2 - h_3 = \Delta h'' = \frac{1}{2}\left(W_3^2 - W_2^2 + U_2^2 - U_3^2\right). \tag{12.6}$$

The stage specific mechanical energy balance requires (Fig. 12.5)

$$\begin{aligned} l_{\mathrm{m}} &= H_1 - H_3 \\ &= (h_1 - h_2) - (h_3 - h_2) + \frac{1}{2}\left(V_1^2 - V_3^2\right). \end{aligned} \tag{12.7}$$

Inserting (12.5) and (12.6) into (12.7) yields

$$l_{\mathrm{m}} = \frac{1}{2}\left[\left(V_2^2 - V_3^2\right) + \left(W_3^2 - W_2^2\right) + \left(U_2^2 - U_3^2\right)\right]. \tag{12.8}$$

Equation (12.8), known as the *Euler turbine equation*, indicates that the stage work can be expressed simply in terms of absolute, relative, and rotational kinetic energies. This equation is equally applicable to turbine stages that *generate* shaft power and to compressor stages that *consume* shaft power. In the case of a turbine stage, the sign of the specific mechanical energy l_{m} is negative, which indicates that energy is removed from the system (power generation). In compressor cases, it is positive because energy is added to the system (power consumption). Before proceeding with velocity diagrams, it is of interest to evaluate the individual kinetic energy differences in (12.8). If we wish to design a turbine or a compressor stage with a high specific mechanical energy l_{m} for a particular rotational speed, then

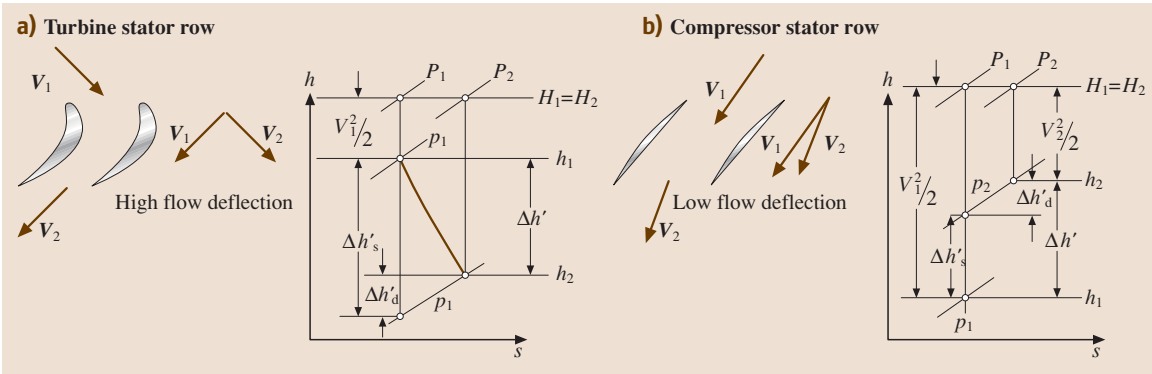


Fig. 12.3a,b Expansion and compression process through a turbine (**a**) and a compressor (**b**) stator row

we have two options:

1. We increase the *flow deflection*, which leads to an increase in $\left(V_2^2 - V_3^2\right)$.
2. We increase the radial difference that leads to a larger $\left(U_2^2 - U_3^2\right)$.

While option 1 is used in axial stages, option 2 is primarily applied to radial stages. These quantities are the characteristics of a stage velocity diagram at the corresponding radial section.

Using the trigonometric relation with the *angle convention* from the velocity diagram in Figs. 12.5 and 12.6, we determine the velocity components and vector relations from

$$
\begin{aligned}
&V_{m2} = W_{m2}\,, \quad V_{m3} = W_{m3}\,,\\
&\boldsymbol{W}_2 = e_1(V_{u2} - U_2) + e_2 V_{m2}\,,\\
&\boldsymbol{W}_3 = -e_1(V_{u3} - U_3) + e_2 V_{m3}\,.
\end{aligned}
\tag{12.9}
$$

In (12.9) V_m, W_m and V_u, W_u are the meridional and circumferential components of the absolute and relative velocities, respectively. The corresponding kinetic energy contributions are determined from

$$
\begin{aligned}
W_2^2 &= \left(V_{u2}^2 + V_{m2}^2\right) + U_2^2 - 2V_{u2}U_2\\
&= V_2^2 + U_2^2 - 2V_{u2}U_2\,,\\
W_3^2 &= V_{u3}^2 + U_3^2 + 2V_{u3}U_3 + V_{m3}^2\,,\\
W_3^2 &= V_3^2 + U_3^2 + 2V_{u3}U_3\,.
\end{aligned}
\tag{12.10}
$$

Incorporating (12.9) and (12.10) into (12.8) yields the *stage specific* work

$$
l_m = U_2 V_{u2} + U_3 V_{u3}\,. \tag{12.11}
$$

Equation (12.11) is valid for axial, radial, and mixed flow turbines and compressors. A similar relation is obtained from the scalar product of moment of momentum and the angular velocity. In conjunction with the equation of moment of momentum, one finds in the

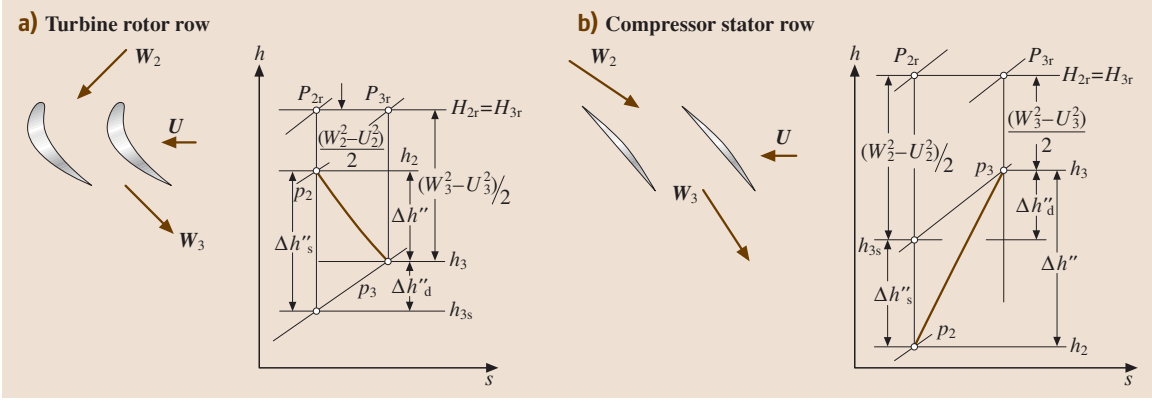


Fig. 12.4a,b Expansion and compression process through a turbine (**a**) and a compressor (**b**) rotor row

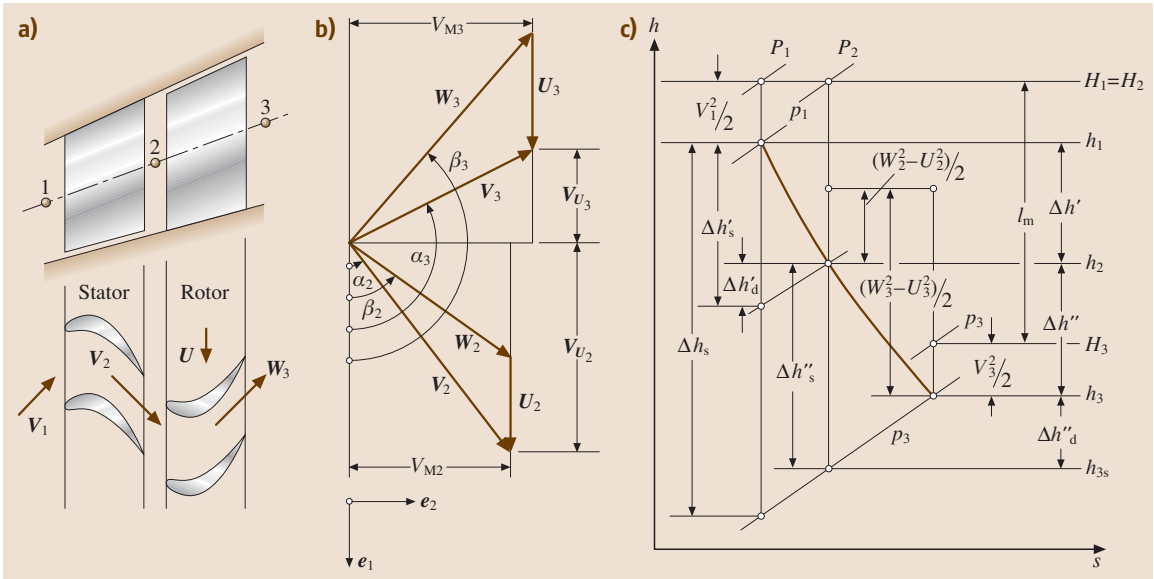


Fig. 12.5a–c A compressor stage (a) with the velocity diagram (b) and the expansion process (c). The direction of the unit vector $\boldsymbol{e}_1$ is identical with the rotational direction

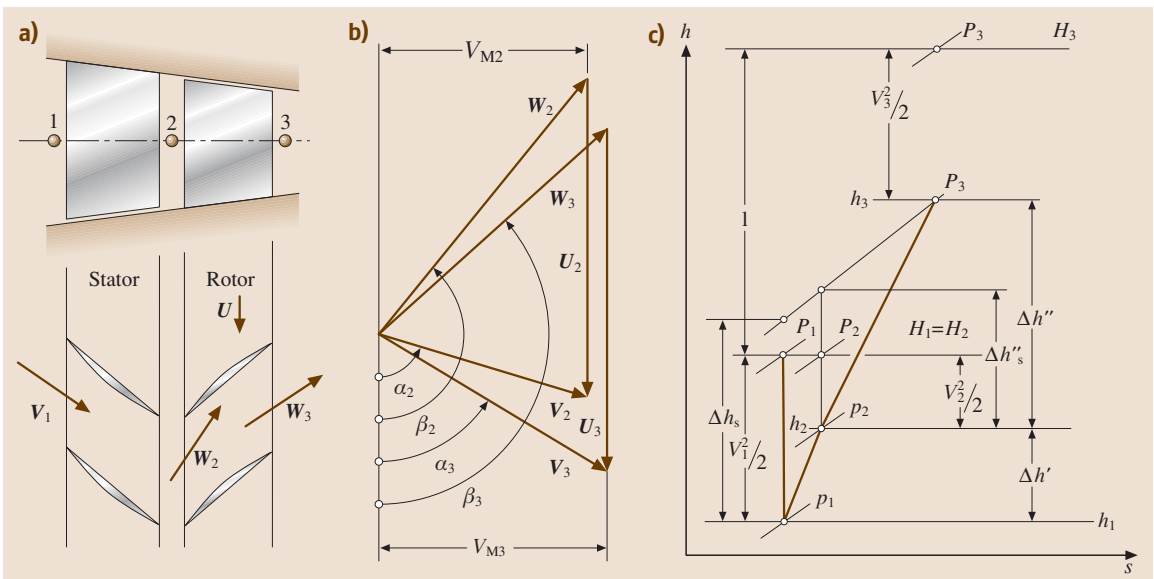


Fig. 12.6a–c A compressor stage (a) with the velocity diagram (b) and the expansion process (c). The direction of the unit vector $\boldsymbol{e}_1$ is identical with the rotational direction

literature $l_m = U_2 V_{u2} - U_3 V_{u3}$. In order to avoid confusion that may arise from different signs, it should be pointed out that the negative sign is the result of the formal derivation of the conservation law of moment of momentum. This negative sign implies that V_{u1} and V_{u2} point in the same direction. The unified angle convention introduced in Figs. 12.1 and 12.2, however, takes the actual direction of the velocity components with regard to a predefined coordinate system.

12.1.4 Dimensionless Stage Parameters

Equation (12.11) exhibits the direct relation between the specific stage work l_m and the kinetic energies. The velocities from which these kinetics energies are built can be taken from the corresponding stage velocity diagram. The objective of this chapter is to introduce dimensionless stage parameters that completely determine the stage velocity diagram. These stage parameters exhibit unified relations for compressor and turbine stages, respectively.

Starting from a turbine or compressor stage with constant mean diameter and axial components (Fig. 12.7) we define the dimensionless stage parameters that describe the stage velocity diagram of a *normal stage* introduced by *Traupel* [12.3]. A normal stage is encountered within the high-pressure (HP) part of multistage turbines or compressor components and is characterized by $U_3 = U_2$, $V_3 = V_1$, $V_{m1} = V_{m3}$, and $\alpha_1 = \alpha_3$. The similarity of the velocity diagrams allows the same blade profile to be used throughout the HP turbine or compressor, thus significantly reducing manufacturing costs. We define the stage flow coefficient ϕ as the ratio of the meridional velocity component and the circumferential component. For this particular case, the meridional component is identical with the axial component

$$\phi = \frac{V_{m3}}{U_3} \,. \tag{12.12}$$

The stage flow coefficient ϕ in (12.12) is a characteristic of the mass flow behavior through the stage. The *stage load coefficient* λ is defined as the ratio of the specific stage mechanical energy l_m and the exit circumferential kinetic energy U_3^2. This coefficient directly relates the flow deflection given by the velocity diagram with the specific stage mechanical energy

$$\lambda = \frac{l_m}{U_3^2} \,. \tag{12.13}$$

The stage load coefficient λ in (12.13) describes the work capability of the stage. It is also a measure for the stage loading. The *stage enthalpy coefficient* Ψ represents the ratio of the isentropic stage mechanical energy and the exit circumferential kinetic energy U_3^2

$$\Psi = \frac{l_s}{U_3^2} \,. \tag{12.14}$$

The stage enthalpy coefficient represents the stage isentropic enthalpy difference within the stage. Furthermore, we define the *stage degree of reaction* r which is the ratio of the static enthalpy difference used in the rotor row divided by the static enthalpy difference used in the entire stage

$$r = \frac{\Delta h''}{\Delta h'' + \Delta h'} \,. \tag{12.15}$$

The degree of reaction r indicates the portion of energy transferred in the rotor blading. Using (12.5) and (12.6), we arrive at

$$r = \frac{\Delta h''}{\Delta h' + \Delta h''} = \frac{W_3^2 - W_2^2 + U_2^2 - U_3^2}{W_3^2 - W_2^2 + U_2^2 - U_3^2 + V_2^2 - V_1^2} \,. \tag{12.16}$$

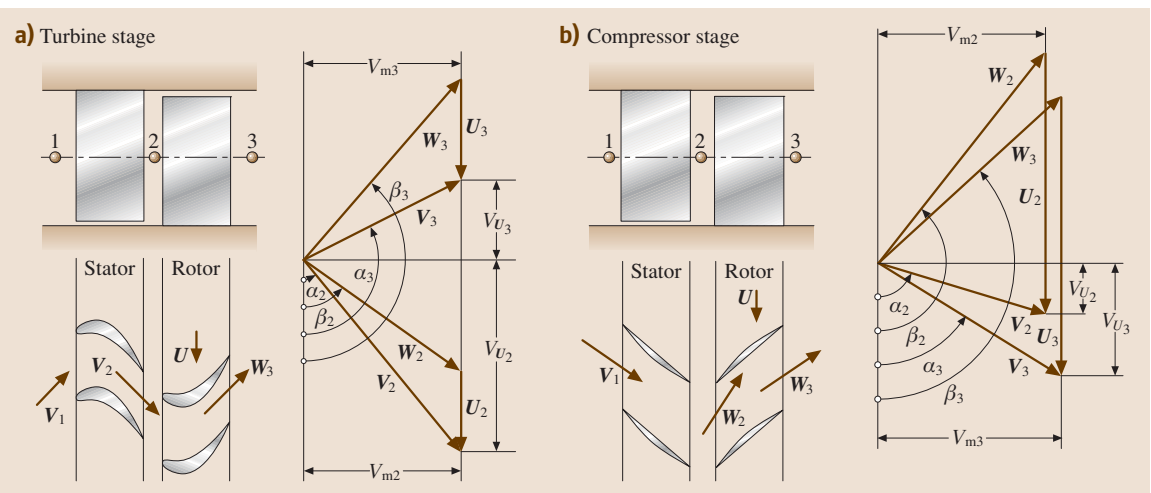


Fig. 12.7a,b Turbine (a) and compressor (b) stages with corresponding velocity diagrams

Since for the stage type under consideration, $V_1 = V_3$ and $U_2 = U_3$, (12.16) can be simplified as

$$r = \frac{W_3^2 - W_2^2}{W_3^2 - W_2^2 + V_2^2 - V_3^2} \,. \tag{12.17}$$

The velocity vectors and the corresponding kinetic energies are determined from the stage velocity diagram in connection with the angle and direction convention as follows

$$\begin{aligned} V_2 &= e_1(W_{u2} + U_2) + e_2 W_{m2} \,, \\ V_2^2 &= (W_{u2} + U_2)^2 + W_{m2}^2 \,, \\ V_3 &= e_1(W_{u3} + U_3) + e_2 W_{m3} \,, \\ V_3^2 &= (W_{u3} + U_3)^2 + W_{m3}^2 \,, \end{aligned} \tag{12.18}$$

since $U_2 = U_3 = U$,

$$V_2^2 - V_3^2 = W_2^2 - W_3^2 + 2UW_{u2} + 2UW_{u3} \,. \tag{12.19}$$

Using (12.18) and (12.19), (12.17) gives

$$r = \frac{W_3^2 - W_2^2}{2U(W_{u2} + W_{u3})} = \frac{W_{u3}^2 - W_{u2}^2}{2U(W_{u2} + W_{u3})} \,. \tag{12.20}$$

Rearranging (12.20) yields the final relationship for the particular stage we introduced above

$$r = \frac{1}{2}\frac{W_{u3} - W_{u2}}{U} \,. \tag{12.21}$$

12.1.5 Relation Between Degree of Reaction and Blade Height for a Normal Stage Using Simple Radial Equilibrium

In axial flow compressors or turbines, the working fluid has a rotational and translational motion. The rotating fluid is subjected to centrifugal forces that must be balanced by the pressure gradient in order to maintain the radial equilibrium. Consider an infinitesimal sector of an annulus with unit depth containing the fluid element which is rotating with tangential velocity V_u.

The centrifugal force acting on the element is shown in Fig. 12.8. Since the fluid element is in radial equilibrium, the centrifugal force is obtained from

$$\mathrm{d}F = \mathrm{d}m\frac{V_u^2}{R} \tag{12.22}$$

with $\mathrm{d}m = \rho R\,\mathrm{d}R\,\mathrm{d}\phi$. The centrifugal force is kept in balance by the pressure forces

$$\frac{\mathrm{d}\rho}{\mathrm{d}R} = p\frac{\mathrm{d}p}{\mathrm{d}\rho}\frac{V_u^2}{R} \,. \tag{12.23}$$

This result can also be obtained by decomposing the Euler equation of motion [12.1, Chap. 3, Eq. (3.46)] for inviscid flows into its three components in a cylindrical coordinate system. The Euler equation is expressed as

$$\boldsymbol{V} \cdot \nabla \boldsymbol{V} = -\frac{1}{\rho}\nabla p \,. \tag{12.24}$$

In the radial direction

$$V_r\frac{\partial V_r}{\partial R} + V_u\frac{\partial V_r}{R\partial\varphi} + V_z\frac{\partial V_r}{\partial z} - \frac{V_u^2}{R} = -\frac{1}{\rho}\frac{\partial p}{\partial R} \,. \tag{12.25}$$

The assumptions needed to arrive at (12.23) are

$$\frac{\partial V_r}{\partial R} \simeq 0 \,, \text{ axial symmetric: } \frac{\partial V_r}{\partial \varphi} = 0 \,, \quad \frac{\partial V_r}{\partial z} \simeq 0 \,. \tag{12.26}$$

With these assumptions, (12.24) yields

$$\frac{1}{\rho}\frac{\partial p}{\partial R} = \frac{V_u^2}{R} \,. \tag{12.27}$$

Equation (12.27) is identical with (12.23). Calculation of a static pressure gradient requires additional information from the total pressure relation. For this purpose, we apply the Bernoulli equation neglecting the gravitational term

$$P = p + \frac{1}{2}\rho V^2 = p + \frac{1}{2}\rho\left(V_u^2 + V_{ax}^2\right) \,. \tag{12.28}$$

Using (12.28), the change in radial direction is

$$\frac{\mathrm{d}p_0}{\mathrm{d}R} = \frac{\mathrm{d}p}{\mathrm{d}R} + \rho V_u\frac{\mathrm{d}V_u}{\mathrm{d}R} + \rho V_{ax}\frac{\mathrm{d}V_{ax}}{\mathrm{d}R} \,. \tag{12.29}$$

If the stagnation or total pressure $P = p_0 = \text{const}$ and $V_{ax} = \text{const}$, (12.29) yields

$$\frac{\mathrm{d}p}{\mathrm{d}R} + \rho V_u\frac{\mathrm{d}V_u}{\mathrm{d}R} = 0 \,, \quad \text{or} \quad \frac{\mathrm{d}p}{\mathrm{d}R} = -\rho V_u\frac{\mathrm{d}V_u}{\mathrm{d}R} \,. \tag{12.30}$$

Equating (12.30) and (12.23) results in

$$V_u\frac{\mathrm{d}V_u}{\mathrm{d}R} + \frac{V_u^2}{R} = 0 \tag{12.31}$$

or

$$\frac{\mathrm{d}V_u}{V_u} + \frac{\mathrm{d}R}{R} = 0 \,. \tag{12.32}$$

The integration of (12.32) leads to $V_u R = \text{const}$. This type of flow is called free vortex flow and fulfills the requirement to be potential flow, $\nabla \times V = 0$. We use

this relation to rearrange the specific stage mechanical energy

$$l_m = U_2 V_{u2} + U_3 V_{u3} = \omega(R_2 V_{u2} + R_3 V_{u3}) \,. \quad (12.33)$$

At station 2 the swirl is $R_2 V_{u2} = \text{const} = K_2$; likewise at station 3 the swirl is $R_3 V_{u3} = K_3$. Since $\omega = \text{const}$, the specific stage mechanical energy is constant

$$l_m = (K_2 + K_3)\omega = \text{const} \,. \quad (12.34)$$

Equation (12.34) implies that, for a stage with constant spanwise meridional components and constant total pressure from hub to tip, the specific stage mechanical energy is constant over the entire blade height. To express the degree of reaction in the spanwise direction, we replace the enthalpy differences in (12.15) by pressure differences. For this purpose we apply the first law for an adiabatic process through stator and rotor blades expressed in terms of $\Delta h'' = \bar{v}'' \Delta p''$ and $\Delta h' = \bar{v}' \Delta p'$. This leads to

$$r = \frac{\bar{v}'' \Delta p''}{\bar{v}'' \Delta p'' + \bar{v}' \Delta p'} = \frac{\Delta p''}{\Delta p'' + \frac{\bar{v}'}{\bar{v}''} \Delta p'} \simeq \frac{p_2 - p_3}{p_1 - p_3} \,. \quad (12.35)$$

In (12.35), the ratio of specific volumes was approximated as $\bar{v}'/\bar{v}'' \simeq 1$. This approximation is admissible for low Mach number ranges. Considering $R_2 V_{u2} = \text{const}$, the integration of (12.23) for station 1 from an arbitrary diameter R to the mean diameter R_m yields

$$(p_1 - p_{m1}) = \frac{\rho}{2} (V_{um})_1^2 \left(1 - \frac{R_m^2}{R^2}\right)_1 \,. \quad (12.36)$$

At station 2 we have

$$(p_2 - p_{m2}) = \frac{\rho}{2} (V_{um})_2^2 \left(1 - \frac{R_m^2}{R^2}\right)_2 \quad (12.37)$$

and finally, at station 3 we arrive at

$$(p_3 - p_{m3}) = \frac{\rho}{2} (V_{um})_3^2 \left(1 - \frac{R_m^2}{R^2}\right)_3 \,, \quad (12.38)$$

with $(R_m)_1 = (R_m)_2 = (R_m)_3$ and $V_{um3} = V_{um1}$. Introducing (12.36)–(12.38) into (12.35), we finally arrive at a simple relationship for the degree of reaction

$$\frac{1 - r}{1 - r_m} - \frac{R_m^2}{R^2} \,. \quad (12.39)$$

From a turbine design point of view, it is of interest to estimate the degree of reaction at the hub and tip radius by inserting the corresponding radii into (12.39). As a result, we find

$$\frac{1 - r_h}{1 - r_m} = \left(\frac{R_m}{R_h}\right)^2 \,, \quad \frac{1 - r_t}{1 - r_m} = \left(\frac{R_m}{R_t}\right)^2 \,. \quad (12.40)$$

Equation (12.40) represents a simple radial equilibrium condition which allows the calculation of the inlet flow angle in a radial direction by integrating (12.32)

$$V_u R = \text{const} \,, \quad R = \frac{\text{const}}{V_u} \,. \quad (12.41)$$

This leads to the determination of the inlet flow angle in a spanwise direction as

$$\frac{R_m}{R} = \frac{\cot \alpha_1}{\cot \alpha_m} \,. \quad (12.42)$$

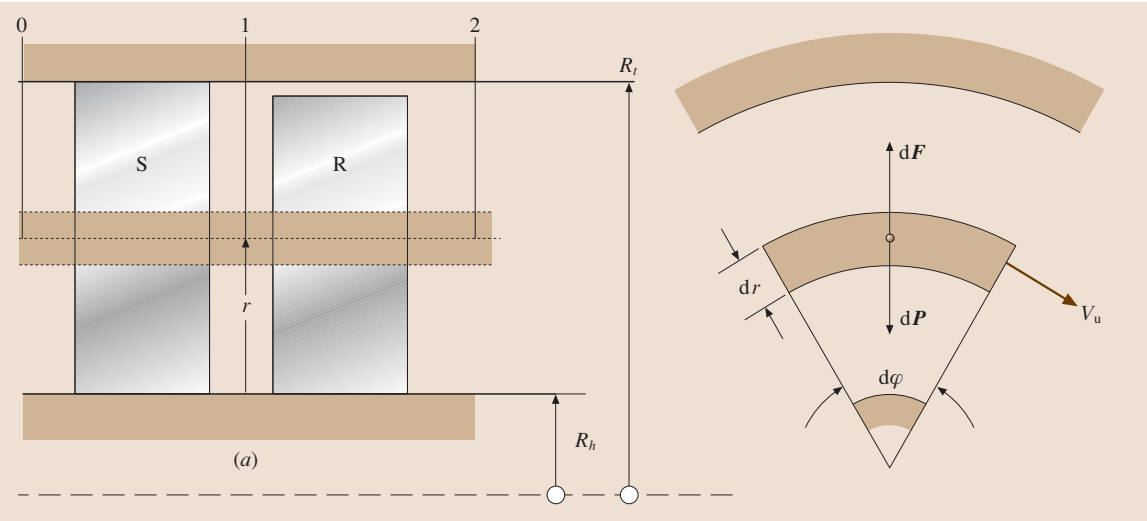


Fig. 12.8 Explanation for simple radial equilibrium

12.1.6 Effect of Degree of Reaction on the Stage Configuration

The distribution of degree of reaction can also be obtained simply by using the velocity ratio for r. If, for example, the degree of reaction at the mean diameter is set at $r = 50\%$, (12.40) immediately calculates r at the hub and tip for the simple radial equilibrium condition $V_u R = \text{const}$, as presented above. It should be mentioned that for a turbine a negative degree of reaction at the hub may lead to flow separation and is not desired. Likewise, for a compressor, r should not exceed the value of 100%. The value of r has major design consequences. For turbine blades with $r = 0$, as shown in Figs. 12.9a and 12.10, the flow is deflected in the rotor and exit velocity vectors have the same magnitude but opposite directions. The entire stage static enthalpy difference is partially converted within the stator row. Note that the flow channel cross section remains constant. For $r = 0.5$, shown in Fig. 12.9b, a fully symmetric blade configuration is established. Figure 12.9c shows a turbine stage with $r > 0.5$. In this case, the flow deflection inside the rotor row is much greater than the one inside the stator row. Figure 12.10 shows the flow deflection within a high-speed rotor cascade. In the past, mainly two types of stages were common designs in steam turbines. The stage with a constant pressure across the rotor blading ($p_2 = p_3$), called an *action stage*, was used frequently. This turbine stage was designed such that the exit absolute velocity vector V_3 was swirl free. It is most appropriate for the design of single-stage turbines and as the last stage of a multistage turbine. The *exit loss*, which corresponds to the kinetic energy of the exiting mass flow, becomes a minimum by using a swirl-free absolute velocity. The stage with $r = 0.5$ is called the *reaction stage*.

12.1.7 Effect of the Stage Load Coefficient on Stage Power

The stage load coefficient λ defined in (12.13) is an important parameter, which describes the capability of the stage to generate/consume shaft power. A turbine stage with low flow deflection, and thus low specific stage load coefficient λ, generates lower specific mechanical energy. To increase the stage mechanical energy l_m, blades with higher flow deflection are used that produce higher stage load coefficient λ. The effect of an increased λ is shown in Fig. 12.11, where three different bladings are plotted.

The top blading with the stage load coefficient $\lambda = 1$ has lower deflection. The middle blading has a moderate flow deflection and moderate $\lambda = 2$, which delivers twice as much stage power as the top blading. Finally, the bottom blading with $\lambda = 3$ delivers three times the stage power as the first one. In the practice of turbine design, among other things, two major parameters must be considered: the specific load coefficients and the stage polytropic efficiencies.

Lower deflection generally yields higher stage polytropic efficiency, but many stages are needed to produce the required turbine power. However, the same turbine

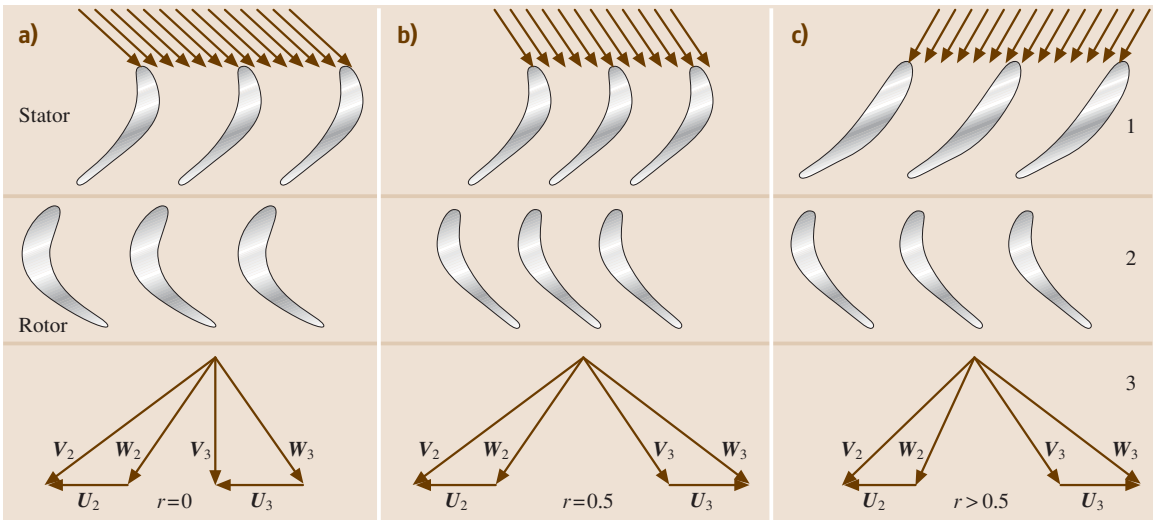


Fig. 12.9a–c Effect of degree of reaction on the stage configuration

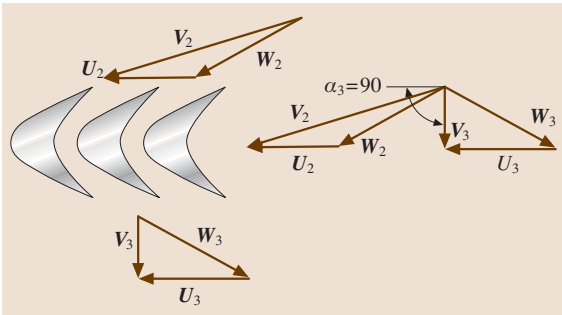


Fig. 12.10 Flow through a high-speed turbine rotor with a degree of reaction $r = 0.0$, note that $\alpha_3 = 90°$ and $|\boldsymbol{W}_2| = |\boldsymbol{W}_3|$

power may be established by a higher stage flow deflection, and thus a higher λ, at the expense of the stage efficiency. Increasing the stage load coefficient has the advantage of significantly reducing the stage number, thus lowering the engine weight and manufacturing cost. In aircraft engine design practice, one of the most critical issues besides the thermal efficiency of the engine is the thrust-to-weight ratio. Reducing the number of stages may lead to a desired thrust-to-weight ratio. While a high turbine stage efficiency has top priority in power generation steam and gas turbine design, the thrust-to-weight ratio is the major parameter for aircraft engine designers.

12.1.8 Unified Description of a Turbomachinery Stage

The following sections treat turbine and compressor stages from a unified standpoint. Axial, mixed flow, and radial flow turbines and compressors follow the same thermodynamic conservation principles. Special treatments are indicated when dealing with aerodynamic behavior and loss mechanisms. While turbine aerodynamics is characterized by negative (favorable) pressure gradient environments, the compression process operates in a positive (adverse) pressure gradient environment. As a consequence, partial or total flow separation may occur on compressor blade surfaces, leading to partial stall or surge. On the other hand, with the exception of some minor local separation bubbles on the suction surface of highly loaded low-pressure turbines blades, a turbine normally operates without major flow separation or breakdown. These two distinctively different aerodynamic behaviors are due to the different pressure gradient environments. Turbine and compressor cascade aerodynamics and losses are extensively treated in [12.1, Chaps. 7 and 16]. In this section, we will first present a set of algebraic equations that describes turbine and compressor stages with constant mean diameter and then extend the approach to general cases where the mean stage diameter changes.

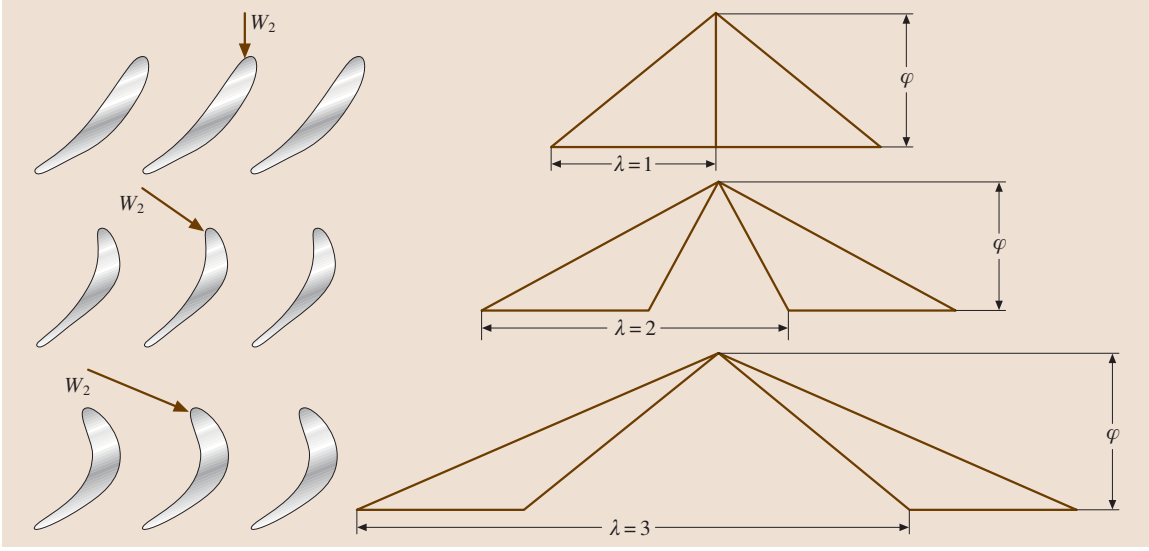


Fig. 12.11 Dimensionless stage velocity diagram to explain the effect of the stage load coefficient λ on flow deflection and blade geometry, $r = 0.5$

Unified Description of Stages with Constant Mean Diameter

For a turbine or compressor stage with constant mean diameter (Fig. 12.7), we present a set of equations that describe the stage by means of dimensionless parameters such as stage flow coefficient ϕ, stage load coefficient λ, degree of reaction r, and the flow angles. From the velocity diagram with the angle definition in Fig. 12.7, we obtain the flow angles

$$\begin{aligned}\cot\alpha_2 &= \frac{U_2 + W_{u2}}{V_{ax}} = \frac{1}{\phi}\left(1 + \frac{W_{u2}}{U}\right)\\ &= \frac{1}{\phi}\left(1 - r + \frac{\lambda}{2}\right),\\ \cot\alpha_3 &= -\frac{W_{u2} - U_2}{V_{ax}} = -\frac{1}{\phi}\left(\frac{W_{u3} - U}{U}\right)\\ &= \frac{1}{\phi}\left(1 - r - \frac{\lambda}{2}\right). \end{aligned} \tag{12.43}$$

The other flow angles can be found similarly, thus we summarize

$$\cot\alpha_2 = \frac{1}{\phi}\left(1 - r + \frac{\lambda}{2}\right), \tag{12.44}$$

$$\cot\alpha_3 = \frac{1}{\phi}\left(1 - \frac{\lambda}{2} - r\right), \tag{12.45}$$

$$\cot\beta_2 = \frac{1}{\phi}\left(\frac{\lambda}{2} - r\right), \tag{12.46}$$

$$\cot\beta_3 = -\frac{1}{\phi}\left(\frac{\lambda}{2} + r\right). \tag{12.47}$$

The stage load coefficient can be calculated from

$$\lambda = \phi(\cot\alpha_2 - \cot\beta_3) - 1\,. \tag{12.48}$$

The velocity diagram of the last stage of a compressor or a turbine differs considerably from the normal stage. As mentioned in the previous section, to minimize the *exit losses*, the last stage usually has an exit flow angle of $\alpha_3 = 90°$. Figure 12.12 compares the velocity diagram of a normal stage with the one in the last stage of the same turbine component. As shown, by changing the exit flow angle to $\alpha_3 = 90°$ a substantial reduction in exit velocity vector V_3, and thus the exit kinetic energy V_3^2, can be achieved. This subject is treated in [12.1, Chap. 7] in detail.

Generalized Dimensionless Stage Parameters

Now we extend the foregoing consideration to compressor and turbine stages where the diameter, circumferential velocities, and meridional velocities are not constant. Examples are axial flow turbine and compressors (Figs. 12.5 and 12.6), radial inflow (centripetal) turbines (Fig. 12.13), and centrifugal compressors (Fig. 12.14).

In the following, we develop a set of unifying equations that describes the above axial turbine and compressor stages, as well as the centripetal turbine and centrifugal compressor stages shown in Figs. 12.13 and 12.14. We introduce new dimensionless parameters

$$\mu = \frac{V_{m2}}{V_{m3}}\,,\quad \nu = \frac{R_2}{R_3} = \frac{U_2}{U_3}\,,\quad \phi = \frac{V_{m3}}{U_3}\,,\quad \lambda = \frac{l_m}{U_3^2}\,,$$
$$r = \frac{\Delta h''}{\Delta h' + \Delta h''}\,, \tag{12.49}$$

with V_m, U from the velocity diagrams and $\Delta h'$ and $\Delta h''$ as the specific static enthalpy difference in the rotor and stator, respectively. The dimensionless parameters μ represents the meridional velocity ratio for the stator and rotor, respectively, ν is the circumferential velocity ratio, ϕ is the stage flow coefficient, λ is the stage load coefficient, and r is the degree of reaction. Introducing these parameters into the equations of continuity, moment of momentum, and the relation for degree of reaction, the stage is completely defined by a set of four

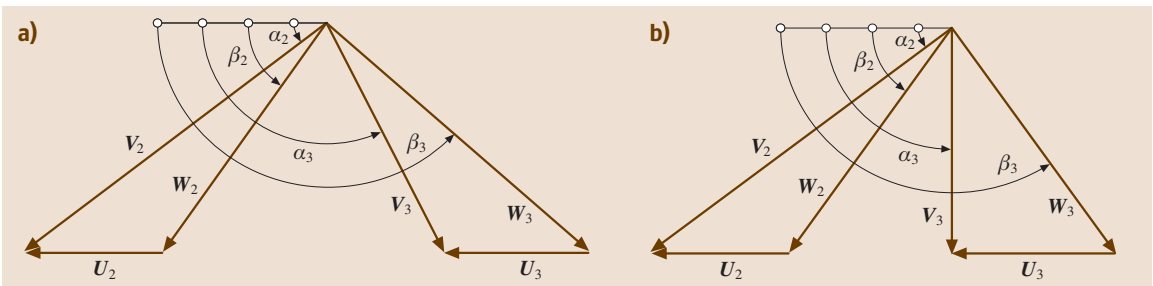


Fig. 12.12a,b Turbine stage velocity diagrams: **(a)** for a normal stage, and **(b)** for the last stage of a multistage turbine

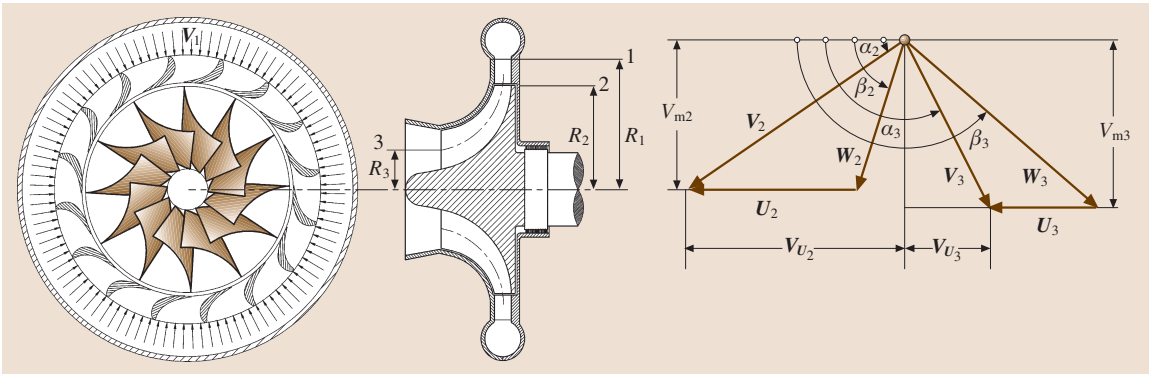


Fig. 12.13 A centripetal turbine stage with cross section and velocity diagram

equations

$$\cot\alpha_2 - \cot\beta_2 = \frac{\nu}{\mu\,\phi}\,, \tag{12.50}$$

$$\cot\alpha_3 - \cot\beta_3 = \frac{1}{\phi}\,, \tag{12.51}$$

$$r = 1 + \frac{\phi^2}{2\lambda}\left[1 + \cot^2\alpha_3 - \mu^2\left(1 + \cot^2\alpha_2\right)\right], \tag{12.52}$$

$$\lambda = \phi(\mu\nu\cot\alpha_2 - \cot\beta_3) - 1\,. \tag{12.53}$$

While (12.50), (12.51) and (12.53) are exact, (12.52) is only an approximation. However, its exact value can be obtained by solving equation system (12.54). The system of (12.50–12.53) contains nine unknown stage parameters. To find a solution, five parameters must be guessed. Appropriate candidates for the first guess are: the diameter ratio, $\nu = R_2/R_3 = U_2/U_3$, the stator and rotor exit angles α_2 and β_3, the exit flow angle α_3, and the stage degree of reaction r. In addition, the stage flow coefficient ϕ can be estimated by implementing the information about the mass flow and using the continuity equation. Likewise, the stage load coefficient λ can be estimated for turbine or compressor stages by employing the information about the compressor pressure ratio or turbine power. Once the five parameters are guessed, the rest of the four parameters are determined by solving the above equation system. In this case, the four parameters calculated fulfill the conservation laws for the particular compressor or turbine blade geometry for which the five stage parameters were guessed. This preliminary estimation of stage parameters, however, is considered the first iteration toward an optimum design. A subsequent loss and efficiency calculation, presented in [12.1, Chap. 8], will clearly show if the guessed parameters were useful or not. In fact, a few iterations are necessary to find the optimum configuration that fulfills the efficiency requirement set by

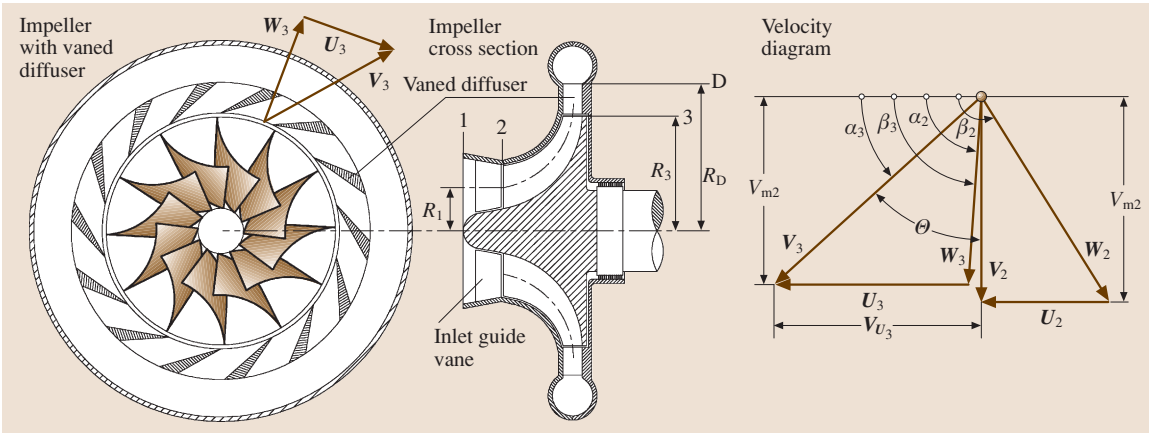


Fig. 12.14 A centrifugal compressor stage with cross section and velocity diagram

the compressor or turbine designer. Equations (12.50–12.53) can be expressed in terms of the flow angles α_2, α_3, β_2, and β_3, which lead to a set of four nonlinear equations

$$\begin{aligned}
&\mu^2\phi^2(1-\nu^2)\cot^2\alpha_2 + 2\mu\nu\,\phi\,\lambda\cot\alpha_2 - \lambda^2 \\
&\quad - 2(1-r)\lambda + (\mu^2-1)\phi^2 = 0\,, \\
&\phi^2(1-\nu^2)\cot^2\alpha_3 + 2\phi\,\lambda\cot\alpha_3 + \lambda^2 \\
&\quad - 2(1-r)\lambda\nu^2 + (\mu^2-1)\phi^2\nu^2 = 0\,, \\
&(1-\nu^2)(\mu\,\phi\cot\beta_2 + \nu)^2 + 2\nu\,\lambda(\phi\,\mu\cot\beta_2 + \nu) - \lambda^2 \\
&\quad - 2(1-r)\lambda + (\mu^2-1)\phi^2 = 0\,, \\
&(1-\nu^2)(\phi\cot\beta_3 + 1)^2 + 2\lambda(\phi\cot\beta_3 + 1) + \lambda^2 \\
&\quad - 2(1-r)\lambda\nu^2 + (\mu^2-1)\phi^2\nu^2 = 0\,.
\end{aligned} \tag{12.54}$$

12.1.9 Special Cases

Equations (12.50–12.54) are equally valid for axial, radial, and mixed flow turbine and compressor stages. Special stages with corresponding dimensionless parameters are described as special cases as listed below.

Case 1: Constant Mean Diameter

In this special case, the diameter remains constant, leading to the circumferential velocity ratio of $\nu = U_2/U_3 = 1$. The meridional velocity ratio is $\mu = V_{m2}/V_{m3} \neq 1$. The flow angles expressed in terms of other dimensionless parameters are

$$\begin{aligned}
\cot\alpha_2 &= \frac{1}{\phi\,\mu}\left[\frac{\lambda}{2} + (1-r) - (\mu^2-1)\frac{\phi^2}{2\lambda}\right], \\
\cot\alpha_3 &= \frac{1}{\phi}\left[-\frac{\lambda}{2} - (1-r) - (\mu^2-1)\frac{\phi}{2\lambda}\right], \\
\cot\beta_2 &= \frac{1}{\mu\,\phi}\left[\frac{\lambda}{2} + (1-r) - (\mu^2-1)\frac{\phi^2}{2\lambda} - 1\right], \\
\cot\beta_3 &= \frac{1}{\phi}\left[-\frac{\lambda}{2} + (1-r) - (\mu^2-1)\frac{\phi}{2\lambda} - 1\right].
\end{aligned} \tag{12.55}$$

The stage load coefficient is calculated from

$$\lambda = \phi(\mu\cot\alpha_2 - \cot\beta_3) - 1 \quad \text{for } \nu = 1 \text{ and } \mu \neq 1\,. \tag{12.56}$$

Case 2: Constant Mean Diameter and Meridional Velocity Ratio

In this special case, the circumferential and meridional velocities are equal, leading to $\nu = U_2/U_3 = 1$, $\mu = V_{m2}/V_{m3} = 1$. The flow angles are calculated from

$$\begin{aligned}
\cot\alpha_2 &= \frac{1}{\phi}\left(\frac{\lambda}{2} - r + 1\right), \\
\cot\alpha_3 &= \frac{1}{\phi}\left(-\frac{\lambda}{2} - r + 1\right).
\end{aligned} \tag{12.57}$$

The stage load coefficient is calculated from

$$\lambda = \phi(\cot\alpha_2 - \cot\beta_3) - 1 \quad \text{for } \nu = 1 \text{ and } \mu = 1\,. \tag{12.58}$$

The generalized stage load coefficient for different μ, ν-cases can be summarized as

$$\begin{aligned}
\lambda &= \phi[\mu\nu\cot\alpha_2 - \cot\beta_3] - 1 \quad \text{for } \nu \neq 1 \text{ and } \mu \neq 1\,, \\
\lambda &= \phi[\mu\cot\alpha_2 - \cot\beta_3] - 1 \quad \text{for } \nu = 1 \text{ and } \mu \neq 1\,, \\
\lambda &= \phi[\nu\cot\alpha_2 - \cot\beta_3] - 1 \quad \text{for } \nu \neq 1 \text{ and } \mu = 1\,, \\
\lambda &= \phi[\cot\alpha_2 - \cot\beta_3] - 1 \quad \text{for } \nu = 1 \text{ and } \mu = 1\,.
\end{aligned} \tag{12.59}$$

12.1.10 Increase of Stage Load Coefficient: Discussion

Following the discussion in Sect. 12.1.3 regarding the increase of the specific stage mechanical energy and the subsequent discussion in Sect. 12.1.8, we proceed with (12.53), where the stage load parameter λ is expressed in terms of μ and ν and the blade angle α_2 and β_3 as

$$\lambda = \phi(\mu\nu\cot\alpha_2 - \cot\beta_3) - 1\,. \tag{12.60}$$

The effect of flow deflection on the stage load coefficient of axial flow turbines was already discussed in Sect. 12.1.8. As we saw, turbine blades can be designed with stage load coefficients λ as high as 3 or more. In turbine blades with high λ and Reynolds numbers $\mathrm{Re} = V_{exit}\,c/\nu > 150\,000$, the governing strong negative pressure gradient prevents major separation from occurring in the flow. However, if the same type of blade operates at lower Reynolds numbers, flow separation that results in a noticeable increase of profile losses may occur. For high-pressure turbines (HPT), the strong favorable pressure gradient within the blade channels prevents major separation from occurring in the flow. However, low-pressure turbine (LPT) blades, particularly those of aircraft gas turbine engines that operate at low Reynolds numbers (cruise conditions up to $\mathrm{Re} = 120\,000$), are subjected to laminar flow separation and turbulent reattachment. While axial turbine blades can be designed with relatively high positive λ,

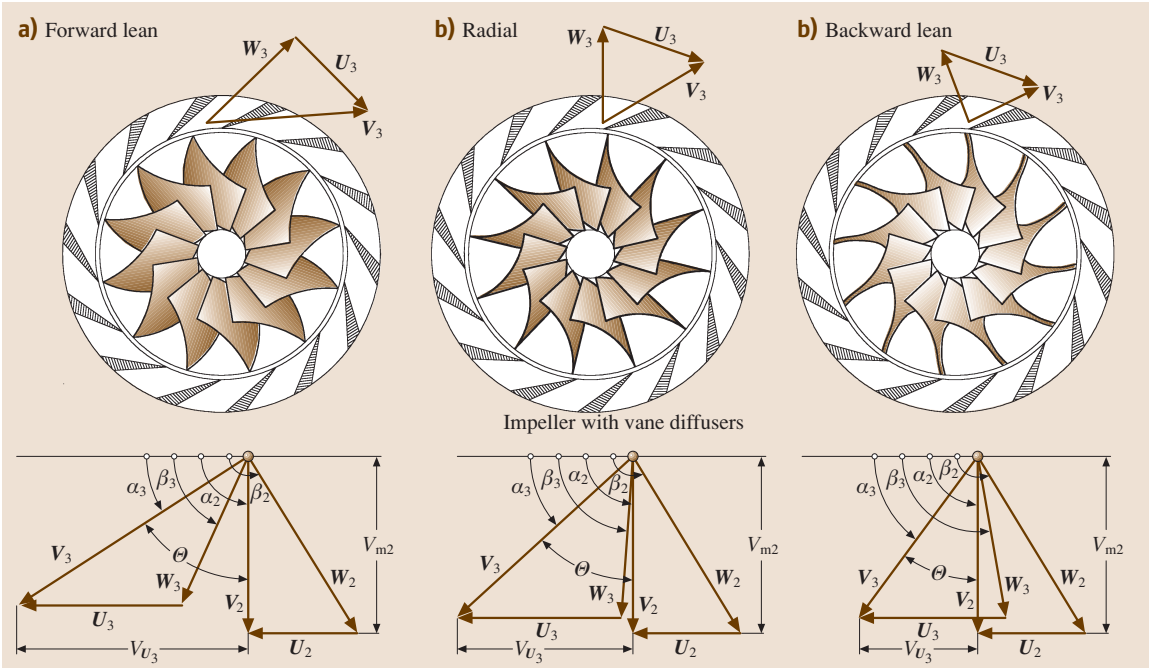


Fig. 12.15a–c Centrifugal compressor stage with velocity diagrams: **(a)** forward lean, **(b)** radial zero lean, and **(c)** backward lean

the flow through axial compressor blade channels is susceptible to flow separation, even at relatively low λ. This is primarily due to a positive pressure gradient in the streamwise direction that causes the boundary layer to separate once a certain deflection limit or a *diffusion factor* (see [12.1, Chap. 16]) has been exceeded. To achieve a higher λ, and thus a higher stage pressure ratio, a smaller diameter ratio $\nu = D_2/D_3 = U_2/U_3$ can be applied. Using a moderate diameter ratio range of $\nu = 0.85 - 0.75$ results in a mixed flow configuration. At a lower range of ν, such as $\nu = 0.75 - 0.4$, a centrifugal compressor configuration is designed.

Figure 12.15 shows, schematically, three centrifugal compressors with three different exit flow angles and the corresponding velocity diagrams. Figure 12.15a shows a centrifugal impeller in which the trailing edge portion is forward leaning with a negative lean angle of $\Delta\beta = \beta_3 - 90° < 0$. The reference configuration Fig. 12.15b shows an impeller with a radial exit flow angle $\beta_3 = 90°$ and thus $\Delta\beta = 0$. Finally, Fig. 12.15c shows an impeller with a backward-leaned trailing edge portion with a positive lean angle of $\Delta\beta = \beta_3 - 90° > 0$. All three

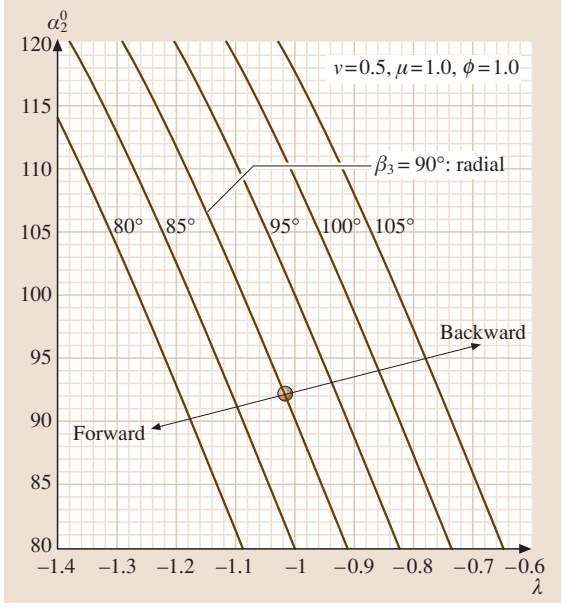


Fig. 12.16 Influence of lean angle on λ: forward lean with $\beta_3 < 90°$, backward lean $\beta_3 > 90°$, and zero lean $\beta_3 = 90°$

impellers have the same diameter ratio ν and the same rotational speed ω. The λ-behavior of these impellers is shown in Fig. 12.16, where the relative exit flow angle is varied in the range of $\beta_3 = 80° - 105°$. As shown, forward lean results in higher deflection Θ, larger ΔV_u, and thus higher negative λ, which is associated with a higher profile loss. Backward lean, however, reduces the flow deflection Θ and ΔV_u. As a result, the stage load coefficient λ is reduced. For the comparison, the radial exit case with $\beta_3 = 90°$ is plotted. In calculating the stage load coefficient λ, the influence of the radius ratio $\nu = R_2/R_3 = U_2/U_3$ on the stage load coefficient becomes clear.

12.2 Gas Turbine Engines: Design and Dynamic Performance

A gas turbine engine is a system that consists of several turbomachinery components and auxiliary subsystems. Air enters the compressor component, which is driven by a turbine component that is placed on the same shaft. Air exits the compressor at a higher pressure and enters the combustion chamber, where the chemical energy of the fuel is converted into thermal energy, producing combustion gas at a temperature that corresponds to the turbine inlet design temperature. The combustion gas expands in the following turbine component, where its total energy is partially converted into shaft work and exit kinetic energy. For power generation gas turbines, the shaft work is the major portion of the above energy forms. It covers the total work required by the compressor component, the bearing frictions, several auxiliary subsystems, and the generator. In aircraft gas turbines, a major portion of the total energy goes toward generation of high exit kinetic energy, which is essential for thrust generation.

Gas turbines are designed for particular applications that determine their design configurations. For power generation purposes, the gas turbine usually has a *single spool*. A spool combines a compressor and a turbine that are connected together via a shaft. Figure 12.17 exhibits a single-spool power generation gas turbine, where a 14-stage compressor shares the same shaft with a three-stage turbine.

While in power generation gas turbine design the power-to-weight ratio does not play an important role, the thrust-to-weight ratio is a primary parameter in designing an aircraft gas turbine. High-performance aircraft gas turbine engines generally have twin-spool or multispool arrangements. The spools are usually rotating at different angular velocities and are connected with each other aerodynamically via air or combustion gas. Figure 12.18 exhibits a typical high-performance twin-spool aircraft gas turbine with a *ducted front fan* as the main thrust generator. Gas turbine engines with power capacities less than 20 MW might have a *split shaft* configuration that consists of a *gas generation spool* and a *power shaft*. While the turbine of the gas generation spool provides the shaft work necessary to drive the compressor, the power shaft produces the net power. In addition to the above design configurations, a variety of engine derivatives can be constructed using a core engine as shown in Fig. 12.19.

12.2.1 Gas Turbine Processes, Steady Design Operation

Starting with the single-spool power generation gas turbine that consists of a multistage compressor, a combustion chamber, and a turbine, the h–s diagram is shown in Fig. 12.20a.

The compression process from 1 to 2 is accomplished by the compressor with a polytropic efficiency η_{pol} that can be accurately calculated using the row-by-row or stage-by-stage methods discussed in [12.6]. The combustion process from 2 to 3 is associated with certain total pressure loss coefficient ζ_{comb}, thus it is not considered isobaric. The expansion process from 3 to 4 causes an entropy increase that is determined by the turbine efficiency. Figure 12.20b shows the h–s diagram for a twin-spool aircraft engine. In contrast to the single-spool engine, the compression process is accomplished by two compressors that are operating at two different angular velocities. Air enters the low-pressure (LP) compressor driven by the LP turbine and is compressed from 1 to 2. Further compression from 2 to 3 occurs in the high-pressure compressor (HP compressor) driven by the HP turbine. After addition of fuel in the combustion chamber, the fist expansion occurs in the HP turbine, whose power exactly matches the sum of the HP compressor power and the power required to compensate bearing frictions. The second expansion in the LP turbine matches the power by the LP compressor, bearing friction, and the auxiliary subsystems. In off-design operation, there is always a dynamic

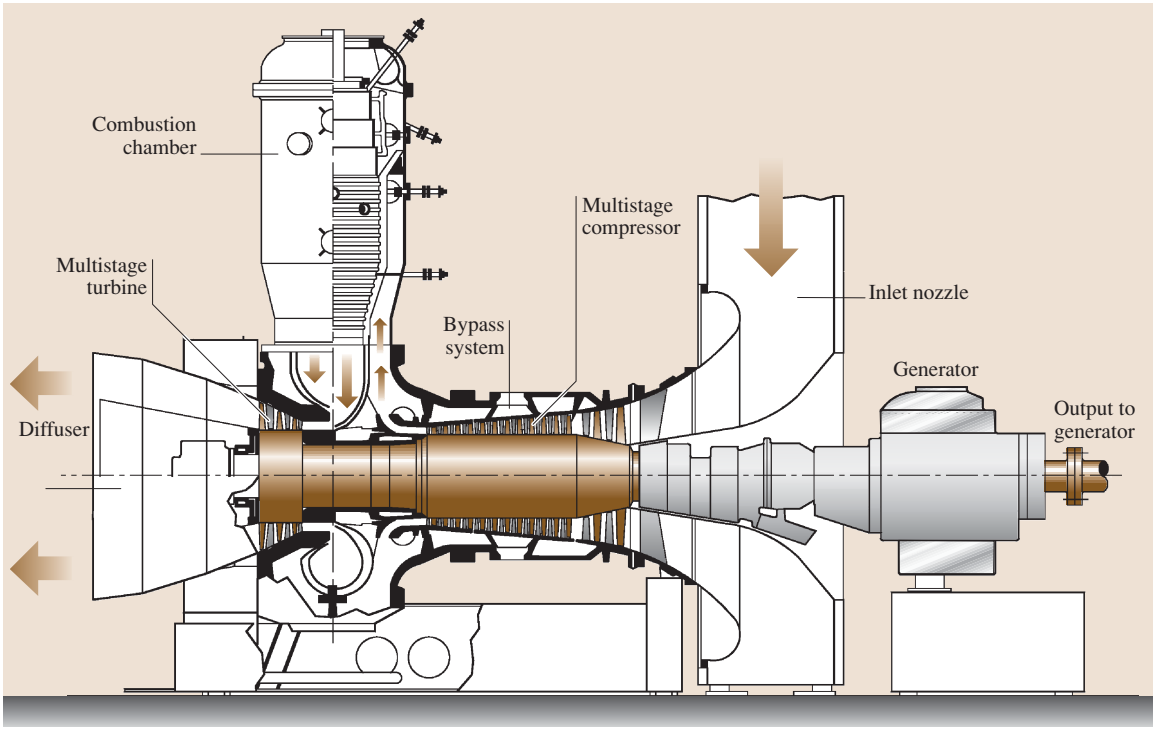


Fig. 12.17 A single-spool power generation gas turbine BBC-GT9

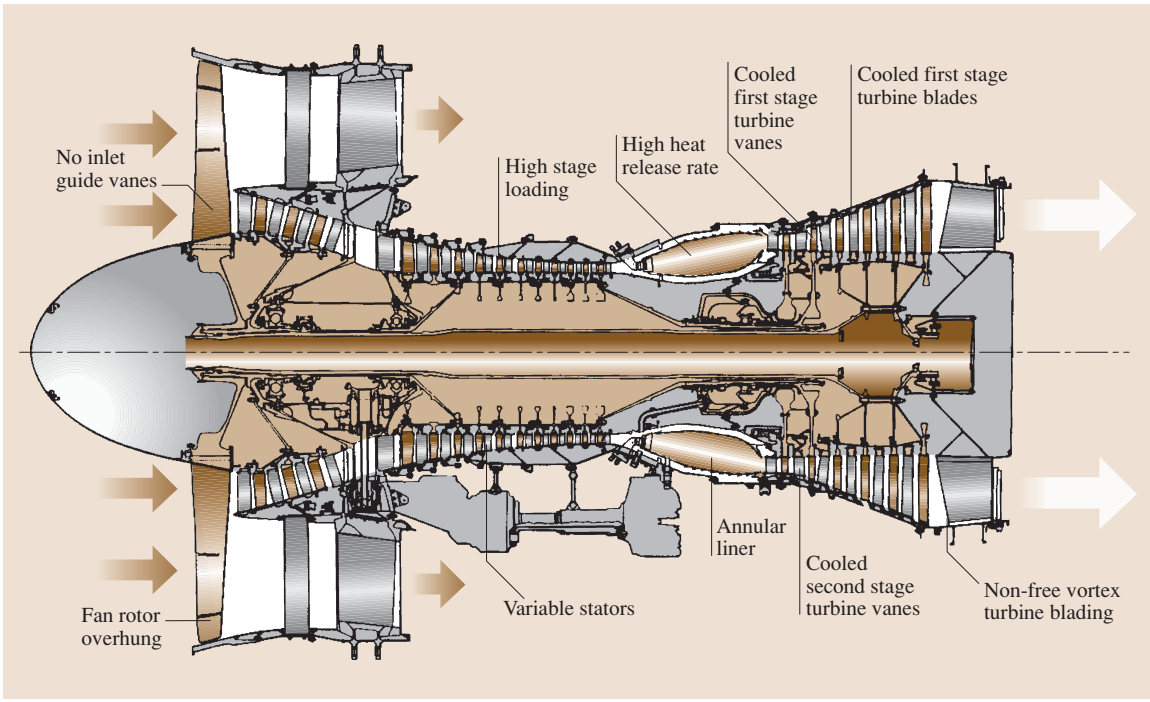


Fig. 12.18 A twin-spool Pratt & Whitney high-bypass-ratio aircraft engine with multistage compressors and turbines

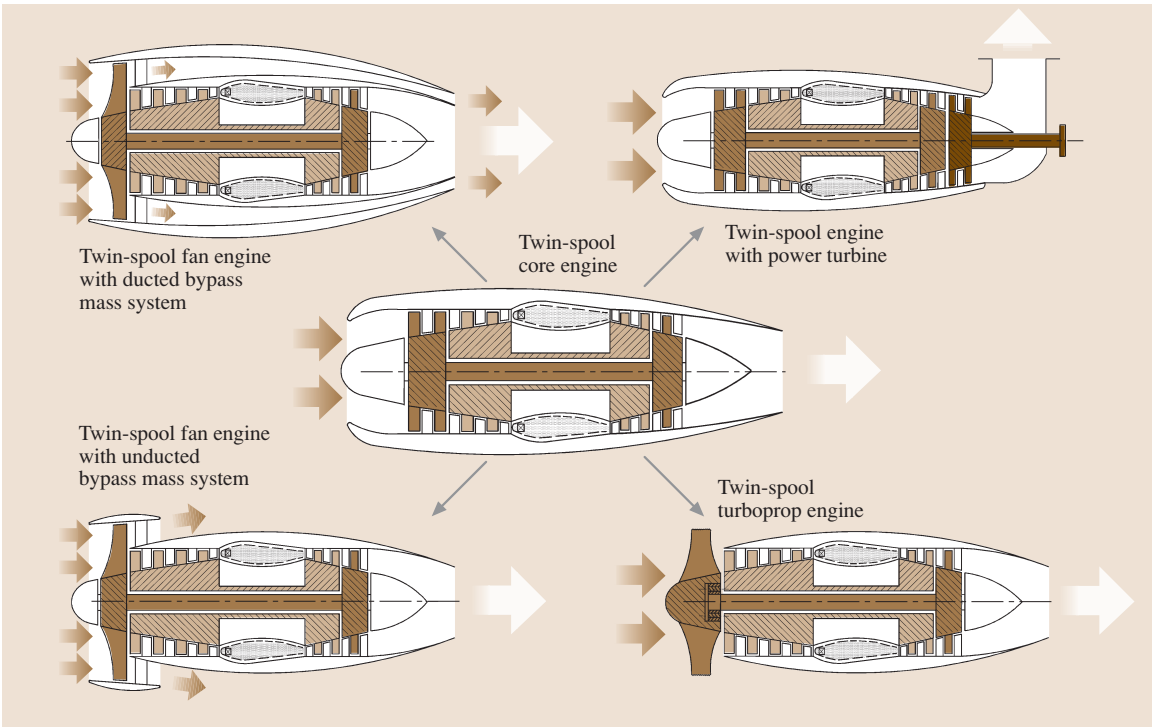


Fig. 12.19 Schematic of a twin-spool core engine with its derivatives

mismatch between the turbine and compressor power that changes the spool rotational speed. This dynamic mismatch is brought to an equilibrium by taking appropriate control measures, as discussed in the following sections. As mentioned briefly, small gas turbines may have a split shaft configuration, as shown in Fig. 12.21. The single-spool gas generator unit provides the power turbine with combustion gas that has the required pressure and temperature to produce the power. As seen in Fig. 12.21a, the specific turbine enthalpy difference is $\Delta h_T \approx \Delta h_C$, leaving the remaining enthalpy difference Δh_{PT} for power generation.

Figure 12.21b shows the h–s diagram of a high-efficiency power generation gas turbine. The schematic cross section of this gas turbine is shown in Fig. 12.25. It consists of a multistage compressor C and a combustion chamber CC1, providing a lean combustion gas that expands in a single-stage reheat turbine (RT). The exhaust gas from the RT enters the second combustion chamber CC2, where the remaining fuel is added to ensure a stoichiometric combustion. It expands in the multistage turbine, which produces the major portion of power. As seen in the following sections, the implementation of the reheat process substantially increases the thermal efficiency of gas turbines. The underlying thermodynamic principles of this concept is the reheat process, which has been very well known in steam turbine design for more than a century. However, in gas turbine design, adding a second combustion chamber to a conventionally designed gas turbine seemed to be associated with unforeseeable problems. Based on design experiences with *compressed air energy storage* (CAES) gas turbines with two combustion chambers, *Brown Boveri* designed and successfully manufactured the first series of power generation gas turbines with a reheat stage and two combustion chambers.

Gas Turbine Process

Accurate prediction of the thermal efficiency of a gas turbine engine requires knowledge of the compressor, combustion chamber, and turbine efficiencies as well as the bearing losses and the losses in auxiliary systems. Furthermore, detailed knowledge of the amount of mass flows with their extraction and injection pressures for cooling the turbine blades and the rotor discs are necessary. In addition, a detailed gas table that accounts for thermodynamic properties of humid air as well as the properties of the combustion gas must be implemented

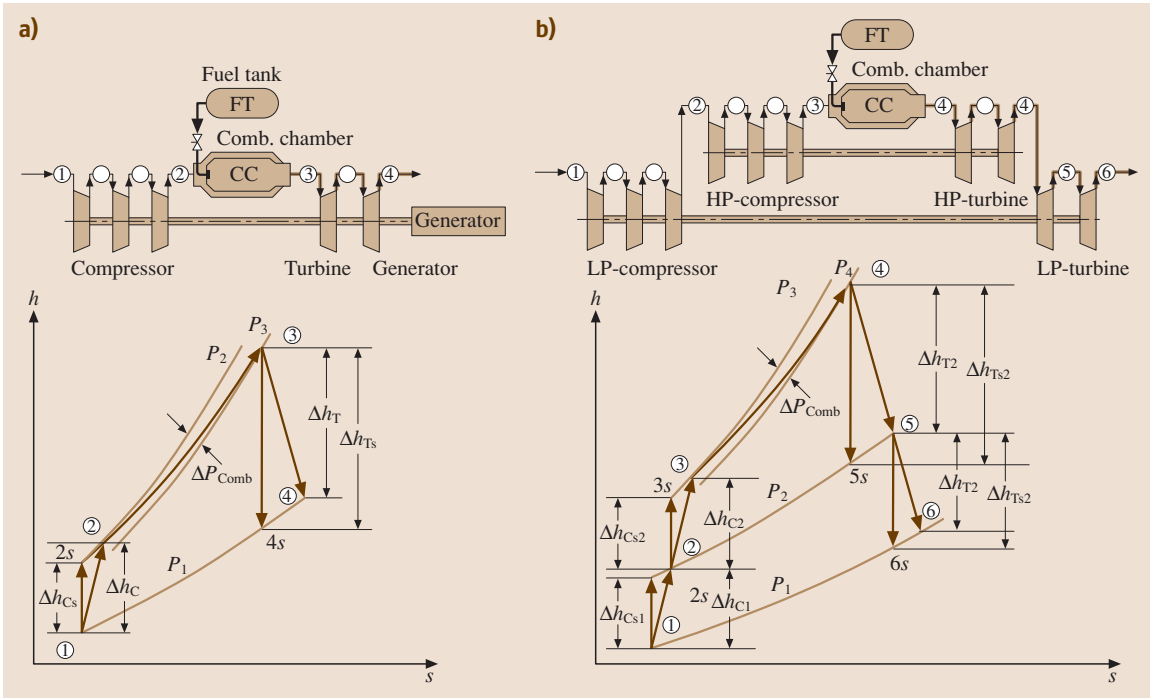


Fig. 12.20a,b *h–s* diagram of **(a)** a single-spool power generation gas turbine and **(b)** a twin-spool aircraft engine

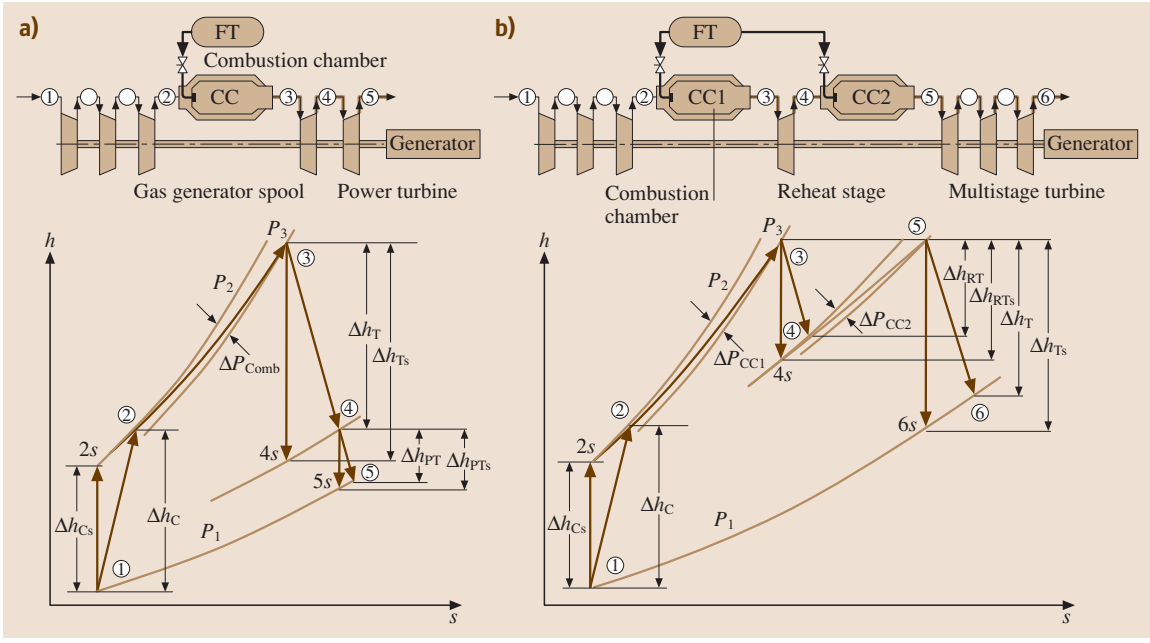


Fig. 12.21a,b *h–s* diagram of **(a)** a single-spool power generation gas turbine with a power shaft and **(b)** a single-spool power generation gas turbine with a reheat stage and two combustion chambers

into the calculation procedure. Assuming that air and combustion gas are calorically perfect gases results in significant errors. Figure 12.22 exhibits a schematic diagram that shows in detail the extraction of different cooling mass flows and their injection locations.

Mass flow through P1 extracted from plenum 3 cools the rotor and does not participate in power generation; mass flows through P2 and P3 cool the second and first turbine stages and remain in the system; and finally, mass flow through P4 reduces the combustion chamber exit temperature before it enters the turbine. At stations 6–8 and 11 humid air is mixed with combustion gas resulting in a local change of water-to-air and fuel-to-air ratios, therefore changing the entire thermodynamic properties including the special gas constant R. In the absence of the above information, reasonable assumptions relative to component efficiencies can be made to qualitatively determine the thermal efficiency and its tendency with regard to parameter variation. In the following section, a simple thermal efficiency calculation procedure is derived that is appropriate for varying different parameters and qualitatively determining their impacts on thermal efficiency.

The gas turbine with its corresponding process is sketched in Fig. 12.23. It consists of a compressor, a recuperator, a combustion chamber, and a turbine. Exhaust gas from the turbine is diverted into the recuperator, heating up the compressed air before entering the combustion chamber. The individual processes are compression, expansion, fuel addition and combustion, and heat exchange in the recuperator. The compressor and turbine enthalpy differences are calculated from

$$h_2 - h_1 = \frac{h_{2s} - h_1}{\eta_c} ,$$
$$h_3 - h_4 = (h_3 - h_{4s})\eta_T . \qquad (12.61)$$

We introduce the following definitions for the recuperator air and gas side (RA, RG), as well as combustion chamber (CC) pressure loss coefficients

$$\zeta_{RA} = \frac{\Delta P_{RA}}{P_2} , \text{ with } \Delta P_{RA} = P_2 - P_5 ,$$
$$\zeta_{RG} = \frac{\Delta P_{RG}}{P_1} , \text{ with } \Delta P_{RG} = P_4 - P_6 ,$$
$$\zeta_{CC} = \frac{\Delta P_{CC}}{P_2} , \text{ with } \Delta P_{CC} = P_5 - P_3 . \qquad (12.62)$$

The thermal efficiency is defined as

$$\eta_{in} = \frac{L_{net}}{\dot{Q}_{in}} = \frac{L_T - L_C}{\dot{Q}_{in}} = \frac{\dot{m}_T l_T - \dot{m}_C l_C}{\dot{Q}_{in}} . \qquad (12.63)$$

The specific net power is calculated from

$$\frac{L_{net}}{\dot{m}_1} = \frac{L_T - L_c}{\dot{m}_1} = \frac{\dot{m}_3 l_T - m_1 l_C}{\dot{m}_1} = (1+\beta) l_T - l_c , \qquad (12.64)$$

with the fuel air ratio $\beta = \dot{m}_f / \dot{m}_1$. Replacing the specific turbine power l_T by the enthalpy difference from

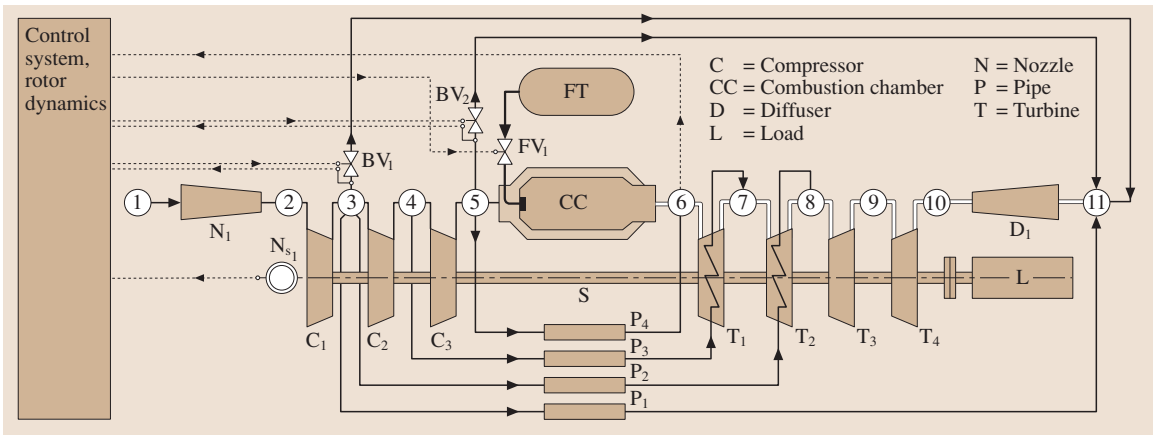


Fig. 12.22 Schematic of a single-spool gas turbine illustrating the mass flow extraction from compressor for different cooling purposes

(12.61), we find

$$\frac{\dot{m}_3}{\dot{m}_1} l_T = (1+\beta)(h_3 - h_4) = \eta_T (1+\beta)(h_3 - h_{4s}) ,$$
$$\frac{\dot{m}_3}{\dot{m}_1} l_T = \eta_T (1+\beta) \bar{c}_{PT} (T_3 - T_{4s}) ,$$
$$\frac{\dot{m}_3}{\dot{m}_1} l_T = \eta_T (1+\beta) \bar{c}_{PT} T_3 \left(1 - \frac{T_{4s}}{T_3}\right) . \qquad (12.65)$$

Equation (12.65) expresses the isentropic enthalpy difference in terms of a product of averaged specific heat at constant pressure and the isentropic temperature difference. The specific heat in (12.65) exhibits an averaged value between the two given temperatures

$$\bar{c}_{PT} = \frac{h_3 - h_{4s}}{T_3 - T_{4s}} . \qquad (12.66)$$

The temperature ratio in (12.65) can be related to the pressure ratio by

$$\frac{T_3}{T_{4s}} = \left(\frac{p_3}{p_4}\right)^{\left(\frac{k-1}{k}\right)_T} = \pi_T^{\left(\frac{k-1}{k}\right)_T} = \pi_T^{m_T} ,$$
$$\text{with } m_T \equiv \left(\frac{k-1}{k}\right)_T . \qquad (12.67)$$

With (12.67), (12.65) becomes

$$\frac{\dot{m}_3}{\dot{m}_1} l_T = \eta_T (1+\beta) \bar{c}_{PT} T_3 \left(1 - \pi_T^{-m_T}\right) . \qquad (12.68)$$

Because of the pressure losses across the combustion chamber, the turbine and compressor pressure ratios are not the same ($\pi_T \neq \pi_c$). Implementing the pressure losses of the combustion chamber and recuperator air side, we find

$$\pi_T = \frac{p_3}{p_4} = \frac{p_2 - \Delta p_{RA} - \Delta p_{CC}}{p_1 + \Delta p_{RA}}$$
$$= \frac{p_2}{p_1} \left(\frac{1 - \zeta_{RA} - \zeta_{CC}}{1 + \zeta_{RA}}\right) = \pi_c \frac{1 - \zeta_{RA} - \zeta_{CC}}{1 + \zeta_{RA}} . \qquad (12.69)$$

We set the fraction on the right-hand side of (12.69) as

$$\epsilon = \frac{1 - \zeta_{RA} - \zeta_{CC}}{1 + \zeta_{RA}} \qquad (12.70)$$

and arrive at

$$\pi_T = \epsilon \pi_c , \text{ for } \epsilon = 0, \zeta_{RA} = \zeta_{CC} = \zeta_{RA} = 0$$
$$\text{and for } \epsilon > 0 \; \zeta_{RA} \neq 0 , \; \zeta_{CC} \neq 0 . \qquad (12.71)$$

For parameter variation, the following values may be used: $\zeta_{RA} \simeq \zeta_{RG} = 2\text{–}8\%$ and $\zeta_{CC} \simeq 5\text{–}10\%$. Following exactly the same procedure defined by (12.64)–(12.71), we find the compressor specific work as

$$l_C = \frac{1}{\eta_C} \bar{c}_{Pc} T_1 \left(\pi_C^{m_C} - 1\right) , \text{ with } m_C = \left(\frac{k-1}{k}\right)_C . \qquad (12.72)$$

Inserting (12.68) and (12.72) into (12.64), we arrive at

$$\eta_{th} = \frac{\eta_T \bar{c}_{PT} T_3 \left[1 - (\epsilon \pi_c)^{-m_T}\right](1+\beta)}{\bar{c}_{PCC} T_1 \left[(1+\beta)\frac{T_3}{T_1} - \frac{T_5}{T_1}\right]}$$
$$- \frac{\frac{1}{\eta_c} \bar{c}_{Pc} T_1 \left(\pi_c^{m_c} - 1\right)}{\bar{c}_{PCC} T_1 \left[(1+\beta)\frac{T_3}{T_1} - \frac{T_5}{T_1}\right]} . \qquad (12.73)$$

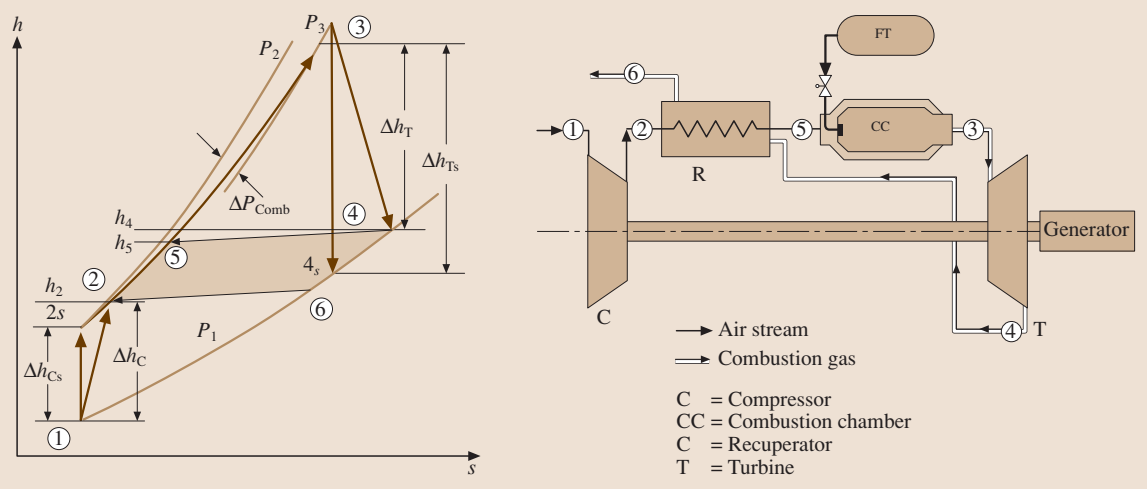


Fig. 12.23 Simple sketch of a gas turbine with recuperator

The turbine inlet temperature T_3 and the environmental temperature T_1, and thus their ratio T_3/T_1, is considered as a known parameter. This parameter can also be used for parametric studies. Therefore it is desirable to express the ratio T_5/T_1 in terms of T_3/T_1. We find this ratio by utilizing the recuperator effectiveness η_R

$$\eta_R = \frac{h_5 - h_2}{h_4 - h_2} \simeq \frac{T_5 - T_2}{T_4 - T_2} \,. \tag{12.74}$$

From the compressor and turbine energy balance in (12.61) we find

$$\begin{aligned} T_2 &= T_1 + (T_{2s} - T_1)\frac{1}{\eta_c} = T_1 + T_1\left(\pi_c^{m_c} - 1\right)\frac{1}{\eta_c} \,, \\ T_4 &= T_3 - (T_3 - T_{4s})\eta_T = T_3 - T_3\left[1 - (\epsilon\pi_c)^{-m_T}\right]\eta_t \,. \end{aligned} \tag{12.75}$$

Equation (12.75) in dimensionless form yields

$$\begin{aligned} \frac{T_2}{T_1} &= 1 + \frac{1}{\eta_c}\left(\pi_c^{m_c} - 1\right) \,, \\ \frac{T_4}{T_1} &= \frac{T_3}{T_1} - \frac{T_3}{T_1}\eta_T\left[1 - (\epsilon\pi_c)^{-m_T}\right] \,. \end{aligned} \tag{12.76}$$

Introducing the temperature ratio $\theta = T_3/T_1$, the temperature ratio T_4/T_1 (12.76) becomes

$$\frac{T_4}{T_1} = \theta\left\{1 - \left[1 - (\epsilon\pi_c)^{-m_T}\right]\eta_T\right\} \,. \tag{12.77}$$

To determine the temperature ratio T_5/T_1, we rearrange (12.74) to obtain

$$\frac{T_5}{T_1} = \eta_R\left(\frac{T_4}{T_1} - \frac{T_2}{T_1}\right) + \frac{T_2}{T_1} \,. \tag{12.78}$$

Using (12.76) and (12.77), (12.78) can be rearranged to

$$\begin{aligned} \frac{T_5}{T_1} = \eta_R\Bigg(&\theta\left\{1 - \left[1 - (\epsilon\pi_c)^{-m_T}\right]\epsilon_T\right\} - 1 \\ &- \frac{1}{\eta_c}\left(\pi_c^{m_c} - 1\right)\Bigg) + 1 + \frac{1}{\eta_c}\left(\pi_c^{m_c} - 1\right) \,. \end{aligned} \tag{12.79}$$

Introducing (12.79) and the definition $\theta = T_3/T_1$ into (12.73), the thermal efficiency equation for a gas turbine with a recuperator is written as

$$\begin{aligned} \eta_{th} = &\Bigg(\bar{c}_{PT}\eta_T\theta\left[1 - (\epsilon\pi_c)^{-m_T}\right](1+\beta) \\ &- \frac{1}{\eta_c}\bar{c}_{Pc}\left(\pi_c^{m_c} - 1\right)\Bigg) \\ &\times\Bigg(\bar{c}_{PCC}\Big\{\theta(1 + \beta - \eta_R) \\ &- \left[1 + \frac{1}{\eta_c}\left(\pi_c^{m_c} - 1\right)\right](1 - \eta_R) \\ &+ \theta\eta_R\eta_T\left[1 - (\epsilon\pi_c)^{-m_T}\right]\Big\}\Bigg)^{-1} \,. \end{aligned} \tag{12.80}$$

From (12.80) special cases are obtained. Setting $\eta_R = 0$ gives the thermal efficiency of a gas turbine without recuperator. The ideal case the of Brayton cycle is obtained by setting all loss coefficients equal to zero, all efficiencies equal to unity, and $\bar{c}_{PC} = \bar{c}_{PCC} = c_{PT} = \text{const}$. Equation (12.80) properly reflects the effects of individual parameters on the thermal efficiency and can be used for preliminary parameter studies. As an example, Fig. 12.8 shows the effect of pressure ratio, the turbine inlet temperature, and the component efficiency on thermal efficiency for two different cases. As Fig. 12.24 shows, for each turbine inlet temperature, there is one optimum pressure ratio. For temperature ratios up to $\theta = 3.5$ pronounced efficiency maxima are visible within a limited π-range. When approaching higher inlet temperature, however, this range widens significantly.

For a gas turbine without recuperator, the thermal efficiency (the solid curves in Fig. 12.24) shows that, for $\theta = 4.0$, increasing the pressure ratio above 15 does not yield a noticeable efficiency increase. However, this requires the compressor to have one or two more stages. The temperature ratio $\theta = 4.0$ corresponds to a turbine inlet temperature of $T_3 = 1200\,\text{K}$ at a compressor inlet temperature of $T_1 = 300\,\text{K}$.

The dashed curves in Fig. 12.24 indicate that tangibly higher thermal efficiencies at a substantially lower pressure ratio can be achieved by utilizing recuperators. This is particularly advantageous for small gas turbines (so called *microturbines*) with power ranging from 50 to 200 kW. The required low maximum pressure ratio can easily be achieved by a single-stage centrifugal compressor. Comparing cases 1 and 2 in Fig. 12.24 shows that thermal efficiency reduces if low-efficiency components are applied.

Improvement of Gas Turbine Thermal Efficiency

The above parameter study indicates that, for a conventional gas turbine with a near-optimum pressure ratio with or without a recuperator, the turbine inlet temperature is the parameter that determines the level of thermal efficiency. For small-size gas turbines, the recuperator is an inherent component of the gas turbine. For large power generation gas turbines, however, this is not a practical option. Using a recuperator in a large gas turbine requires a significantly lower pressure ratio, which results in a large-volume recuperator and turbine. As a result, in order to improve the thermal efficiency of conventional gas turbines, increasing the turbine inlet temperature seems to be the only option left. Considering this fact, in the past three decades,

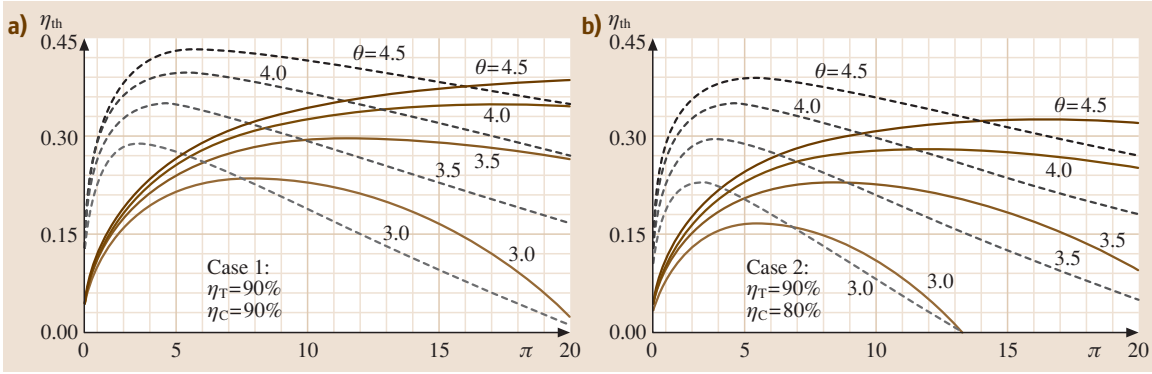


Fig. 12.24a,b Thermal efficiency as a function of pressure ratio with turbine inlet temperature ratio as a parameter for **(a)** a gas turbine with a recuperator (*dashed curves*) and **(b)** without a recuperator (*solid curves*). $\eta_R = 0.75$, $\zeta_{CC} = 0.05$, $\zeta_{RA} = \zeta_{RG} = 0.03$, for case 1 and case 2. In case 2 the turbine efficiency is lowered from 90% to 80%

gas turbine manufacturer have been concentrating their efforts on introducing more sophisticated cooling technologies, which are essential for increasing the turbine inlet temperature of conventional gas turbines.

To improve the thermal efficiency substantially without a significant increase in turbine inlet temperature, the well-known reheat principle as a classical method for thermal efficiency augmentation is applied. Although this standard efficiency improvement method is routinely applied in steam turbine power generation, it did not find its way into aircraft and the power generation gas turbine design. The reason for this was the inherent problem of integrating a second combustion chamber into a conventionally designed gas turbine engine. This issue raised a number of unforeseeable design integrity and operational reliability concerns. ABB (formerly Brown Boveri & Cie) was the first to develop a gas turbine engine with one reheat stage turbine followed by a second combustion chamber and a multistage turbine (Fig. 12.25).

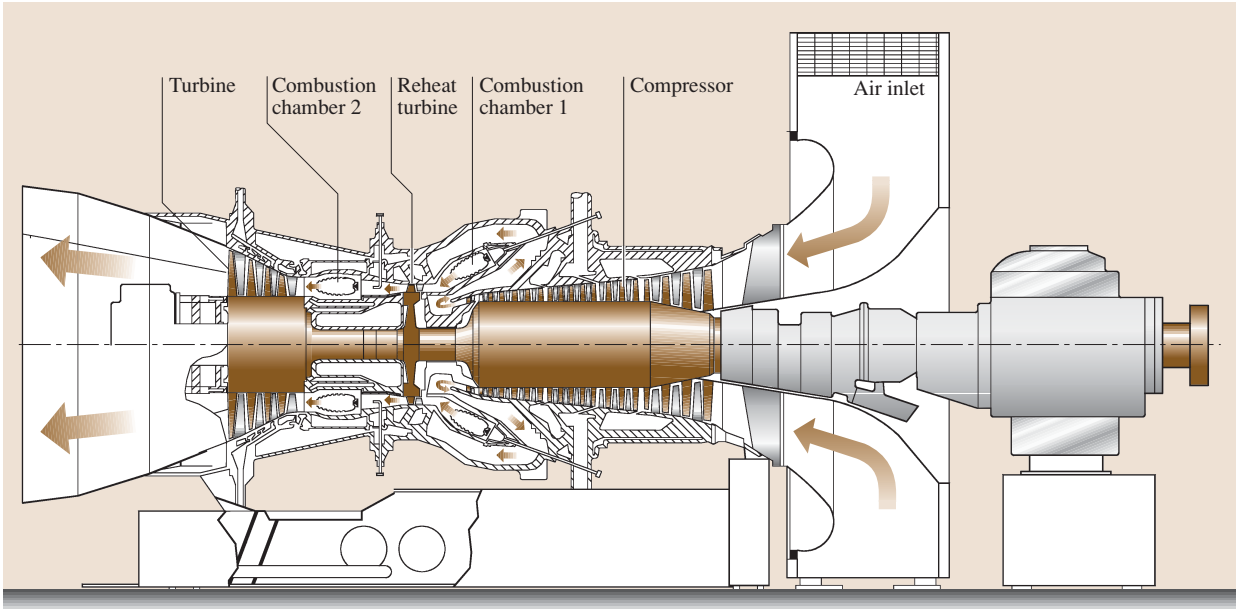


Fig. 12.25 A schematic cross section of the GT-24 gas turbine engine with a single-stage reheat turbine and a second combustion chamber

A comparative study in [12.7] emulates, among others, two conceptually different power generation gas turbine designs utilizing components whose detailed aerodynamic performance characteristics were known. The first is a conventional gas turbine, whereas the second one has a reheat turbine stage and a second combustion chamber resembling the GT-24 [12.8]. Starting from a given environmental condition (pressure, temperature) and a consolidated turbine inlet temperature $T_{3BL} = 1200\,°C$ for both engines (Fig. 12.26a) the thermal efficiency is determined by the compressor pressure ratio and the compressor and turbine polytropic efficiencies η_c and η_T and is plotted in Fig. 12.27a. As curve 1 shows, for the given pressure ratio, which is not identical with the optimum pressure ratio, an efficiency of $\eta_{th} = 35\%$ is calculated. Substantial efficiency improvement is achieved by introducing a single-stage reheat principal, as applied to the GT-24. Details of the process are sketched in Fig. 12.26b with the baseline process as the reference process. The bright dotted area in Fig. 12.26b translates into the efficiency improvement, which in the case of the GT-24 resulted in efficiency improvement of 5.5% above the baseline efficiency. A detailed dynamic engine simulation of the GT-24 gas turbine engine with GETRAN© [12.9] verified a thermal efficiency of $\eta_{th} = 40.5\%$ plotted in Fig. 12.27a, curve 2. This tremendous efficiency improvement was achieved despite the facts that (a) the compressor pressure ratio is much higher than the optimal one for baseline engine and (b) the introduction of a second combustion chamber inherently causes additional total pressure losses. Further calculation showed that introducing a third combustion chamber would only result in a marginal improvement of 1–1.5% thermal efficiency, which does not justify the necessary research and development efforts to integrate a third combustion chamber. The specific work comparison is plotted in Fig. 12.27b, which shows a significant increase in specific work. Additional efficiency improvement requires a technology change. Major improvement can be achieved by using the ultrahigh-efficiency gas turbine (UHEGT) technology [12.8]. This technology eliminates the combustion chambers altogether and places the combustion process inside the stator blade passages (Sect. 12.2.6).

12.2.2 Nonlinear Gas Turbine Dynamic Simulation

The continuous improvement of efficiency and performance of aircraft and power generation gas turbine systems during the past decades has led to engine designs that are subject to extreme load conditions. The engine components operate near their aerodynamic, thermal, and mechanical stress limits. Under these circumstances, any adverse dynamic operation causes excessive aerodynamic, thermal, and subsequent mechanical stresses that may affect the engine safety, and reliability if adequate precautionary actions are not taken. Considering these facts, an accurate predic-

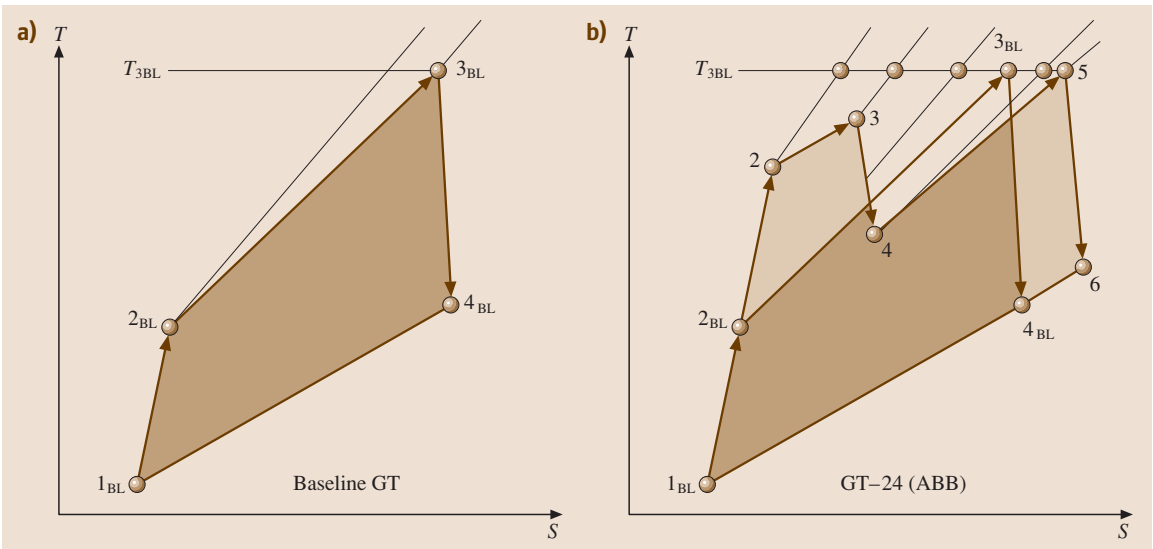


Fig. 12.26a,b Comparison of (a) a conventional baseline gas turbine process (b) with the GT-24 process after [12.7]

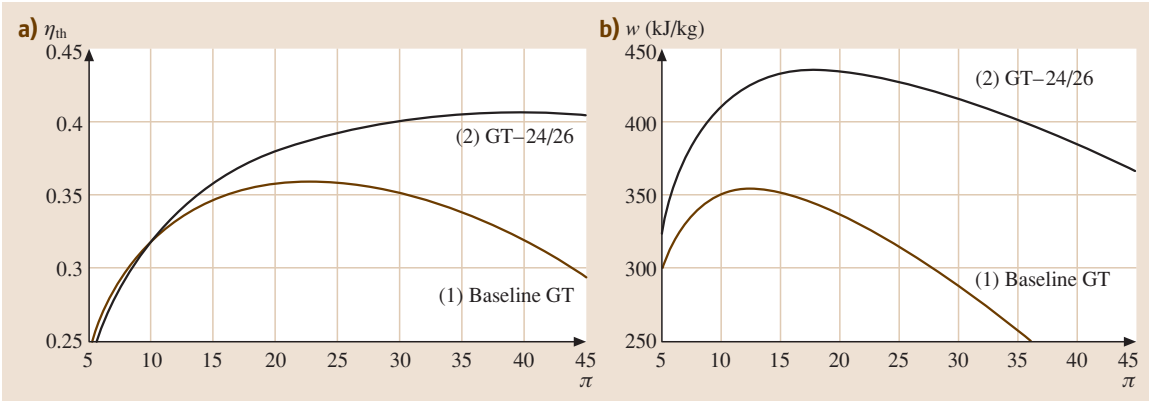


Fig. 12.27a,b Comparison of efficiency and specific work between a conventional baseline gas turbine and GT-24: **(a)** efficiency, **(b)** specific work

tion of the above stresses and their cause is critical at the early stages of design and development of the engine and its components. This section focuses on the simulation of the dynamic behavior of gas turbine engines and their components. The simulation spectrum encompasses single- and multispool gas turbine engines, turbofan engines, and power generation gas turbine engines. The simulation concept is based on a generic modularly structured system configurations. This concept is discussed in detail in [12.6], where the gas turbine components are represented by individual modules described mathematically by systems of differential equations. Based on these and other necessary modules, a generic concept is presented that provides the reader with necessary tools for developing computer codes for simulation of arbitrary engine and plant configurations ranging from single-spool thrust generation to multispool thrust/power generation engines under adverse dynamic operating conditions. It can easily be extended to rocket engines, combined cycles, cogeneration cycles, and steam power plants. In [12.6] a multilevel system simulation treats different degrees of complexity ranging from global adiabatic simulation to a detailed diabatic one. The dynamic behavior of the subject engine is calculated by solving a number of systems of partial differential equations which describe the unsteady behavior of the individual components. Accurate prediction of the dynamic behavior of the engine and the identification of critical parameters by using the method enables the engine designer to take appropriate steps using advanced control systems. The method may also be used to proof the design concept of the new generation of high-performance engines. The modular structure of the concept enables the user to independently develop new components and integrate them into the simulation code. As representative examples, four different case studies are presented that deal with dynamic simulation of a compressed air energy storage gas turbine, different transient cases with single- and multispool thrust, and power generation engines were simulated. The transient cases range from operating with a prescribed fuel schedule, to extreme load changes and generator shut down.

12.2.3 Engine Components, Modular Concept, and Module Identification

A schematic component arrangement and modeling of a twin-spool core engine is shown in Fig. 12.28. The corresponding modules are implemented into the engine modular configuration schematic in Fig. 12.29. Figures 12.30 and 12.31 show the lists of components with their corresponding modular representations and symbols that are described by the method presented in [12.6]. They exhibit the basic components essential for generically configuring any possible aero- and power generation gas turbine engines. These modules are connected with each other with a plenum, which is a coupling component between two or more successive components. As briefly explained in [12.6], the primary function of the plenum is to couple the dynamic information of entering and exiting components such as mass flow, total pressure, total temperature, fuel-to-air ratio, and water-to-air ratio. After entering the plenum a mixing process takes place, where the aforementioned quantities reach their equilibrium values. These values are the same for all outlet components.

A survey of power and thrust generation gas turbine engines has led to the practical conclusion that any arbitrary aircraft or power generation gas turbine engine and its derivatives, regardless of configuration, i. e., number of spools and components, can be generically simulated by arranging the components according to the engine configuration of interest. The nonlinear dynamic method presented in [12.6] is based on this generic, modularly structured concept that simulates the transient behavior of existing and new engines and their derivatives. The modules are identified by their names, shaft number, and inlet and outlet plena. This information is vital for automatically generating the system of differential equations representing individual modules. Modules are then combined into a complete system which corresponds to the engine configuration. Each module is physically described by the conservation laws of thermo-fluid mechanics which result in a system of nonlinear partial differential or algebraic equations. Since an engine consists of a number of components, its modular arrangement leads to a system containing a number of sets of equations. The above concept can be systematically applied to any aircraft or power generation gas turbine engine.

The general application of the modular concept is illustrated in Figs. 12.28 and 12.29. The twin-spool engine shown in Fig. 12.28 exemplifies the modular extension of the single-spool base engine. It consists of two spools with shafts S_1 and S_2, on which the low- and high-pressure components such as compressors and turbines are assembled. The two shafts are coupled by the working media air and combustion gas. They rotate with different speeds, which are transferred to the control system by the sensors Ns_1 and Ns_2. Air enters the inlet diffuser D_1, which is connected with the multistage compressor assembled on S_1, and is decomposed in several compressor stages C_{1i}. The first index (1) refers to the spool number and the second index i marks the number of the compressor stage. After compression in the S_1 compressor stage group, the air enters the second compressor (HP compressor) assembled on the S_2 shaft, which consists of stages C_{21}–C_{25}. In the combustion chamber (CC_1) high-temperature combustion gas is produced by adding the fuel from the tank FT. The gas expands in the high-pressure turbine that consists of stages T_{21}–T_{23}. When exiting from the last stage of the HP turbine, the combustion gas enters the low-pressure turbine, consisting of stages T_{11}–T_{13} and is expanded

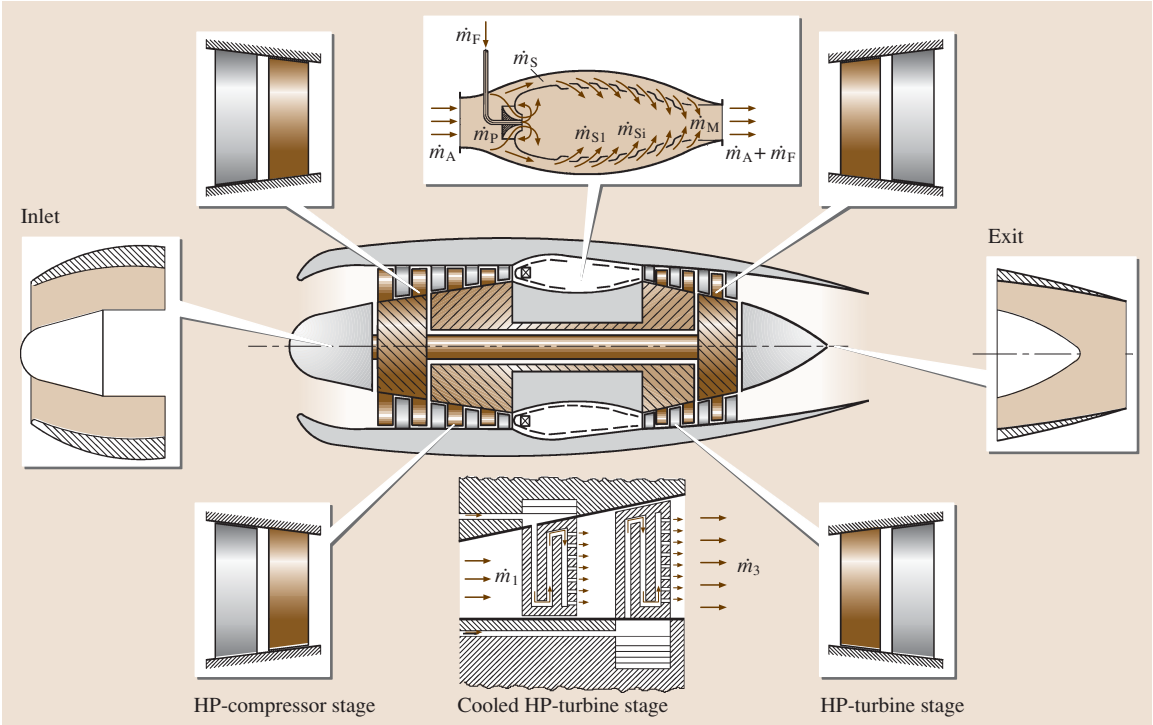


Fig. 12.28 Schematic of a twin-spool core engine, component decomposition

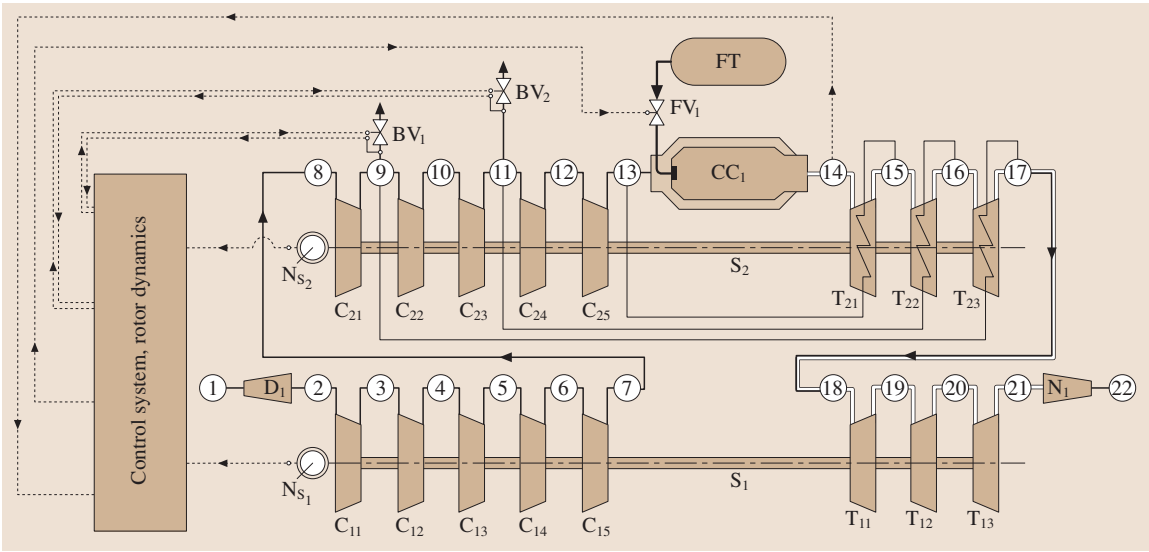


Fig. 12.29 Modular configuration of the engine exhibited in Fig. 12.28

through the exit nozzle. Two bypass valves, BV_1 and BV_2, are connected with the compressor stator blades for surge prevention. The fuel valve, FV_1, is placed between the fuel tank, FT, and the combustion chamber, CC_1. The pipes, P_i, serve for cooling air transport from the compressor to cooled turbines. The compressor stage pressures, the turbine inlet temperature, and the rotor speed are the input signals to the control system, which controls the valve cross sections and the fuel mass flow.

Figure 12.32 shows a more complex example of a three-spool supersonic engine with its modular decomposition. Figure 12.33 exhibits a systematic modular configuration of Fig. 12.32 that is represented by a large system of differential and algebraic equations.

12.2.4 Levels of Gas Turbine Engine Simulations, Cross Coupling

The accuracy of gas turbine dynamic simulation is determined by the level of component modeling. It increases by increasing the level of simulation complexity. Four levels of simulation are introduced:

- The *zeroth simulation level* is applied to simple cases utilizing a fixed system configuration with steady-state component characteristics that are described by algebraic equations, simplified differential equations, and lookup tables and maps. Furthermore, there is no dynamic coupling between the components. Since this simulation level does not account for engine dynamics, it will not be discussed further.
- The *first simulation level* uses the component global performance map only for turbines and compressors. The maps are generated using the row-by-row adiabatic calculation method detailed in [12.6]. The other components such as recuperators, coolers, combustion chambers, pipes, nozzles, and diffusers are simulated according to methods discussed in [12.6]. Primary air, secondary combustion gas, and metal temperature of the combustion chamber are calculated. All modules are coupled with plena, ensuring a dynamic information transfer to all modules involved. Modules are described by algebraic and differential equations.
- The *second simulation level* utilizes adiabatic row-by-row or stage-by-stage calculation for the compressor and turbine modules. For combustion chamber, primary air, secondary combustion gas, and metal temperature are calculated. Dynamic calculations are performed throughout the simulation, where the modules are coupled by plena. Each module is described by differential and algebraic equations.
- The *third simulation level* uses diabatic row-by-row calculation for compressor and turbine modules. This level delivers very detailed diabatic information about the compressor and turbine component dynamic behavior. It utilizes cooled turbine and

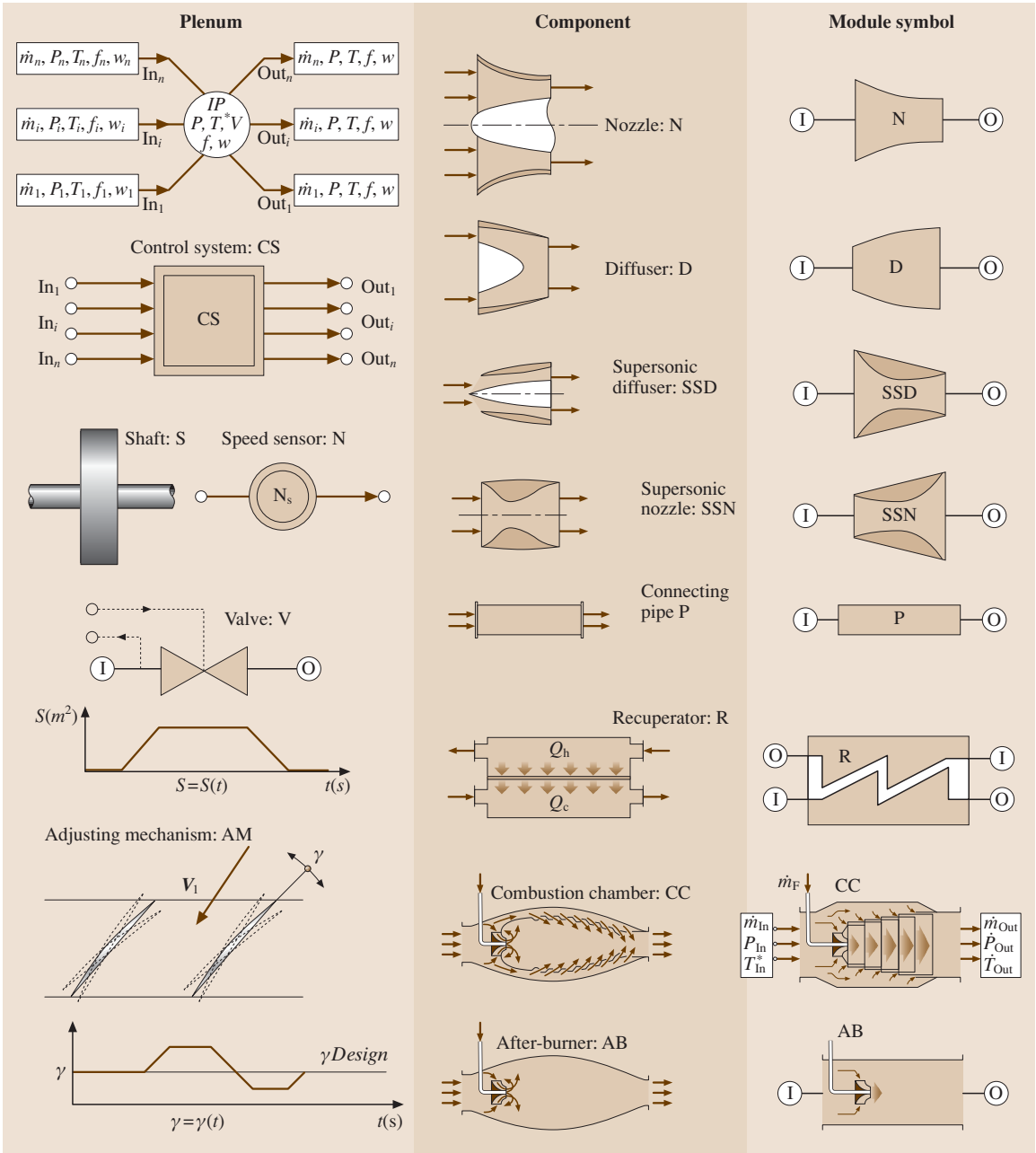


Fig. 12.30 Components, modules, and their symbols: plenum, control system CS, shaft S, with moment of inertia I and the rotational velocity ω, speed sensor N, valve with an arbitrary ramp for closing and opening the cross section s, adjusting mechanism AM for stator blade adjustment, subsonic nozzle N, subsonic diffuser D, supersonic diffuser SSD, supersonic nozzle SSN, recuperator R, combustion chamber CC, and afterburner AB

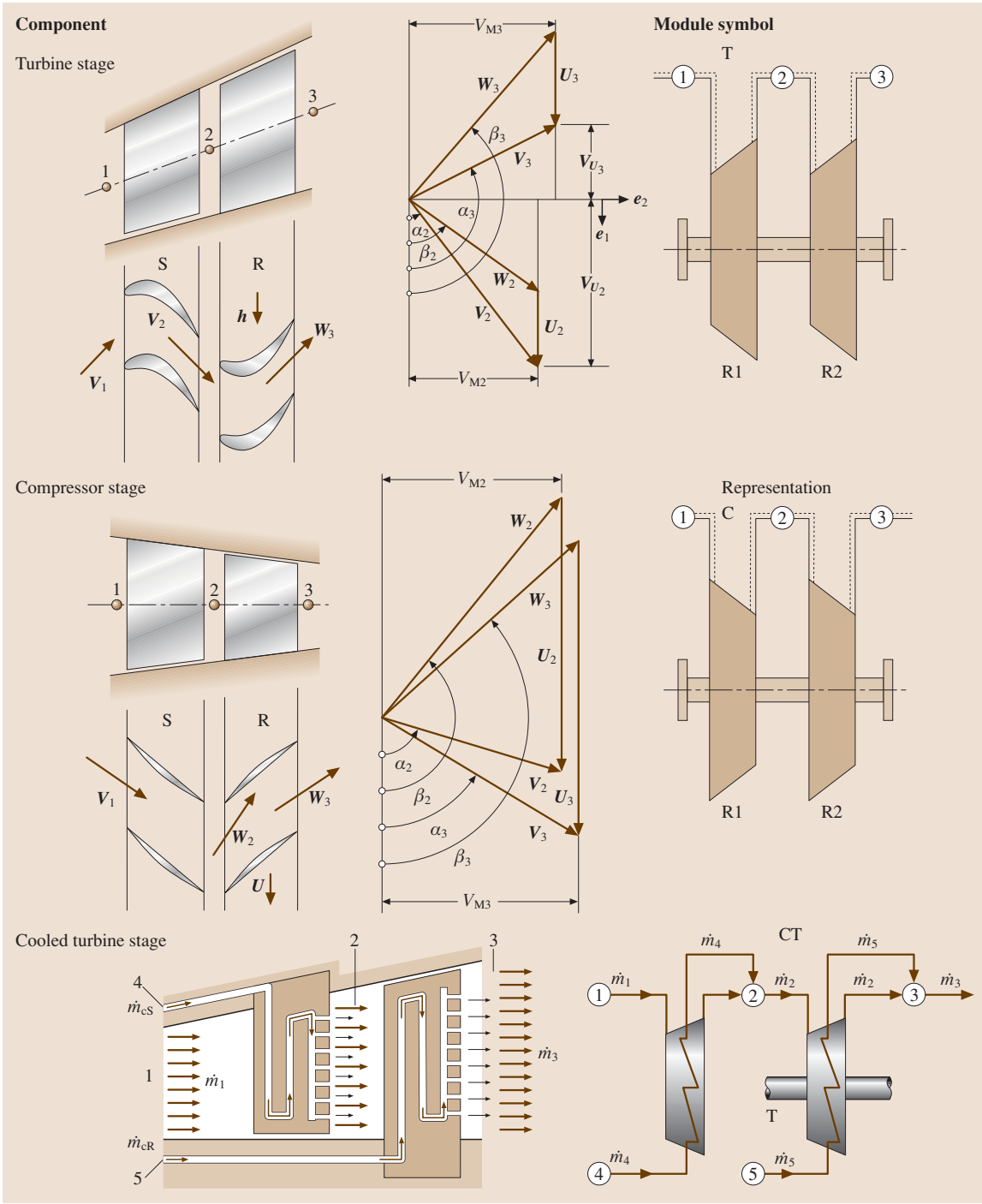


Fig. 12.31 Adiabatic turbine stage with the module T, adiabatic compressor stage with the module C, cooled turbine stage with module CT

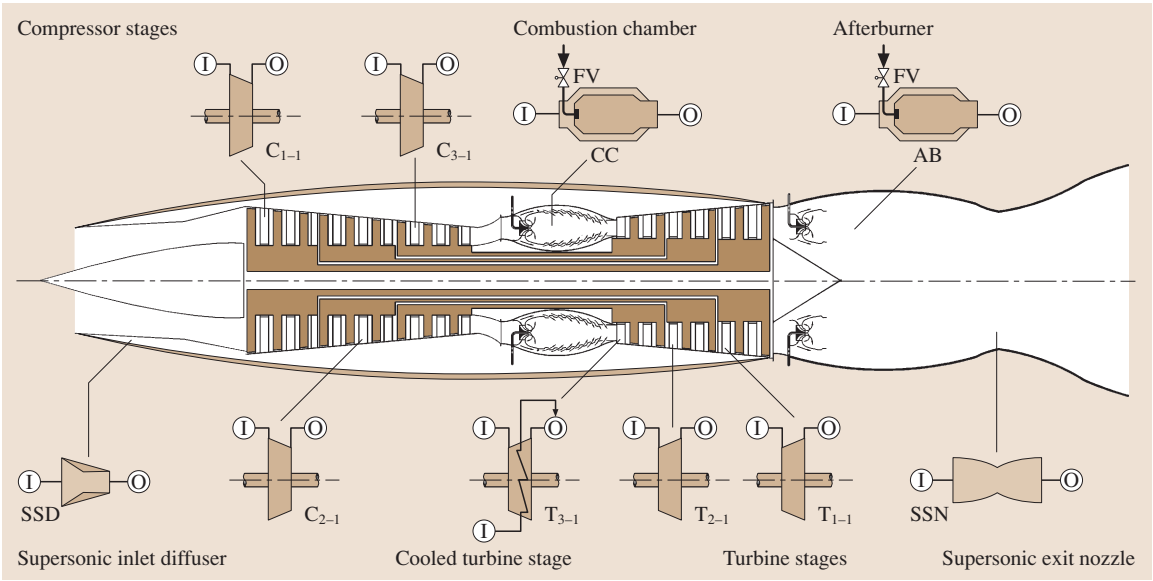


Fig. 12.32 Schematic of a three-spool high-performance core engine, component decomposition

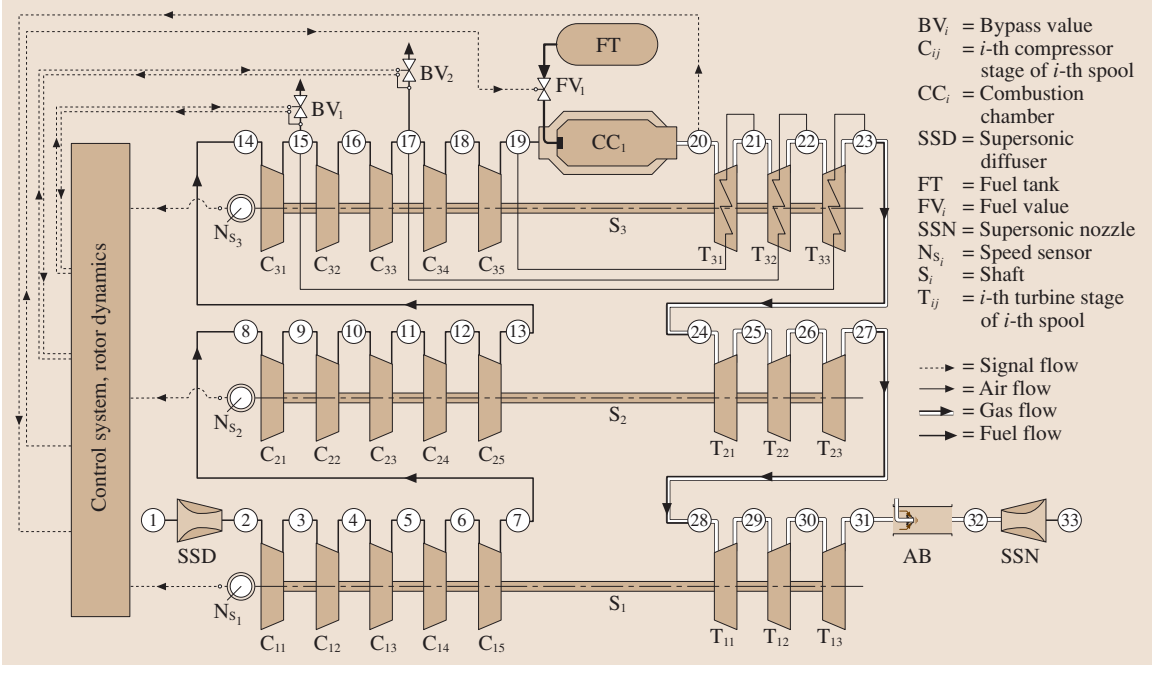


Fig. 12.33 Modular configuration of the three-spool engine shown in Fig. 12.31. The three compressors and turbines are connected aerodynamically. The plena addressing, the spool number, and the component number uniquely identify the set of differential equations that describe the module

compressor stages and simultaneously calculates the blade temperatures. For the combustion chamber, primary air, secondary combustion gas, and metal temperature are calculated. Dynamic calculations are performed throughout the simulation, whereas the modules are coupled by plena. Each module is described by differential and algebraic equations. The details of information delivered by this level and degree of complexity is demonstrated by the following example. The first two stages of a four-stage turbine component of a high-performance gas turbine engine must be cooled. For the first four turbine rows we use the diabatic expansion process that requires three differential equations for describing the primary flow, three differential equations for describing the cooling flow, and one differential equation for describing the blade temperature. This leads from two cooled turbine stages to 28 differential equations.

The generic structure allows to cross-couple levels 1 to 3. For example, we wish to simulate a gas turbine engine with a global compressor performance map, but need to obtain detailed information about turbine blade temperature, which is necessary to calculate the relative expansion between the blades and the casing, then we may use the diabatic calculation method. In this case, we cross-couple the first- and third-level simulation.

12.2.5 Nonlinear Dynamic Simulation Case Studies

Three case studies dealing with three completely different gas turbine systems are presented. Table 12.1 shows the matrix of the cases where the engine types and transient-type simulations are listed. These studies demonstrate the capability of the generic structured method discussed in [12.6] to simulate complex systems dynamically and with high accuracy. The case studies presented in this chapter are related to real-world engine simulation and are intended to provide the reader with an insight into nonlinear engine dynamic simulation. The selected cases ranging from zero-spool, single-shaft power generation to three-spool four-shaft thrust and power generation gas turbine engines provide detailed information about the engine behavior during design and off-design dynamic operation. For each engine configuration the simulation provides aerothermodynamic details of each individual component and its interaction with the other system components. Since the presentation of the complete simulation results of the three cases listed in Table 12.1 would exceed the scope of this chapter, only a few selected plots will be displayed and discussed for each case.

Case Study 1: Compressed Air Energy Storage Gas Turbine

The subject of this case study is a zero-spool, single-shaft compressed air energy storage (CAES) gas turbine [12.1], which is utilized to cover peak electric energy efficiently demand during the day. Continuous increases of fuel costs have motivated the power generation industry to invest in technologies that result in fuel saving. Successful introduction of combined cycle gas turbines (CCGT) has drastically improved the thermal efficiency of steam power plants, which is equivalent to a significant fuel saving. Further saving is achieved by using the excess electrical energy available during the period of low electric energy demand (6–8 h during the night) to compress air into a large storage system. During periods of peak demand, the compressed air is injected into the combustion chambers and mixed with the fuel. After the ignition process is completed, the high-pressure high-temperature gas expands in the turbine, generating electric energy for about 2–4 h. In contrast to a CCGT, the period of operation of a CAES

Table 12.1 Simulation case studies

Tests	Gas turbine type	Transient type
Case 1	CAES: Compressed air energy storage power generation gas turbine engine, zero-spool, single shaft, two turbines, two combustion chambers.	Generator and turbine shut down.
Case 2	Single-spool, single-shaft, power generation gas turbine engine, BBCGT9.	Adverse load changes.
Case 3	Three-pool, four-shaft, thrust and power generation core engine.	Operation with fuel schedule.

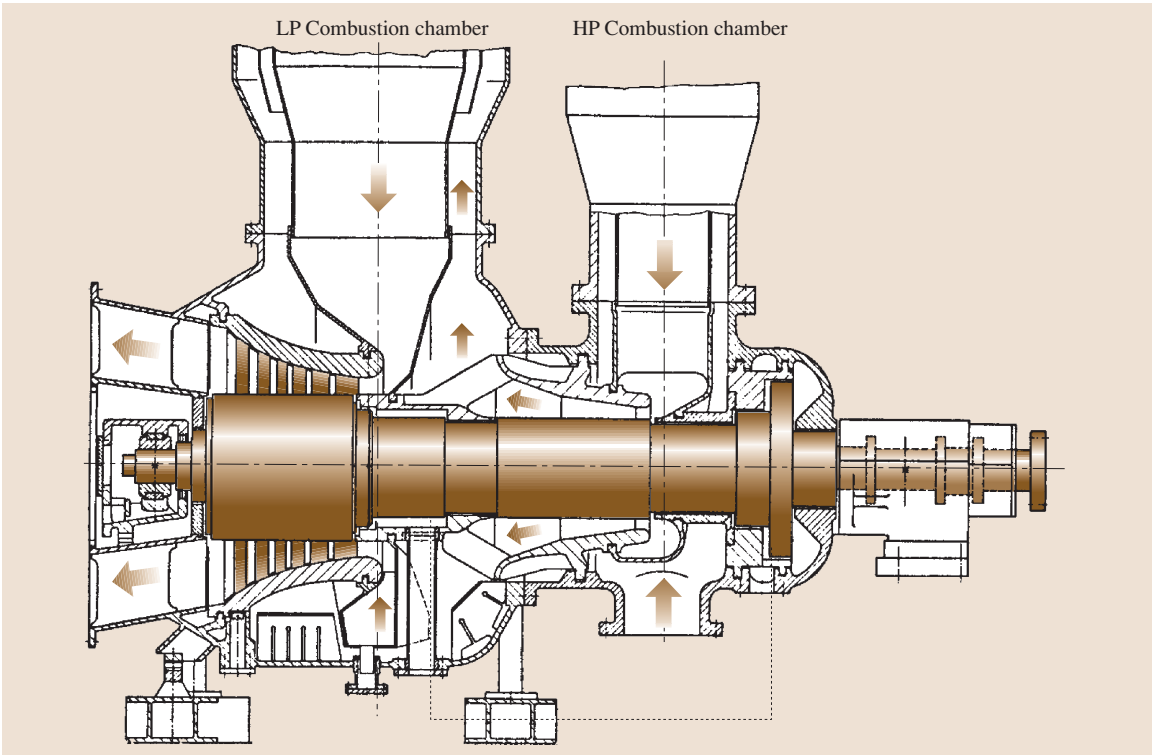


Fig. 12.34 BBC-CAES-Huntorf gas turbine engine after [12.1]

plant is restricted to a few hours per day, resulting in a daily startup followed by a shutdown procedure. This relatively high frequency of startups and shutdowns may cause structural damages resulting in reduced lifetime if the startup and shutdown procedures are not performed properly. The condition for a safe startup procedure is outlined in this study, which helps the engine and control system designer to integrate this into their design procedure. The CAES gas turbine system (Fig. 12.34) with the simulation schematic shown in Fig. 12.35 features a large-volume plenum (8) for storing the compressed air, a high-pressure combustion chamber (HPCC), a high-pressure turbine (HPT), a low-pressure combustion chamber (LPCC), a low-pressure turbine (LPT2), a cold-air preheater with a low- and high-pressure side (LPP and HPP side) and a generator (G). During steady-state turbine operation, cold air from the air-storage facility, plenum 8, passes through the shutdown valve (V_1) to the inlet plenum (1), where it is divided into combustion and cooling-air flows. The addition of fuel in the HPCC causes the combustion air to be heated to the combustion chamber's exit temperature. Immediately upstream of the HPT, the combustor mass flow is mixed with a portion of the cooling-air flow, which has already been preheated in the HPP. As a result, the gas temperature of the turbine mass flow lies below the combustion chamber's exit temperature. After expansion in the HPT, the combustion chamber (LPCC) mass flow is mixed in the LPT inlet plenum (4) with the rest of the preheated cooling-air flow and the sealing-air flow. After expansion in the LPT, the gas gives off some of its heat in the LPP before leaving the gas turbine system.

Figure 12.35 shows how the various components are interconnected. Plenum 8, the air storage facility, is connected via two identical pipes (P6) to two shutdown valves (V_1). During steady-state operation, the blow-off valve (V_2) remains closed, being opened in the event of a disturbance likely to cause rapid shutdown. In such an event, the valve blows off some of the gas, thereby limiting the maximum rotor speed. For the sake of clarity, the preheater (P) has been separated into its air and gas sides, designated by HPP and LPP, respectively.

Simulation of Emergency Shutdown. Starting from a steady operating point, a generator trip with rapid

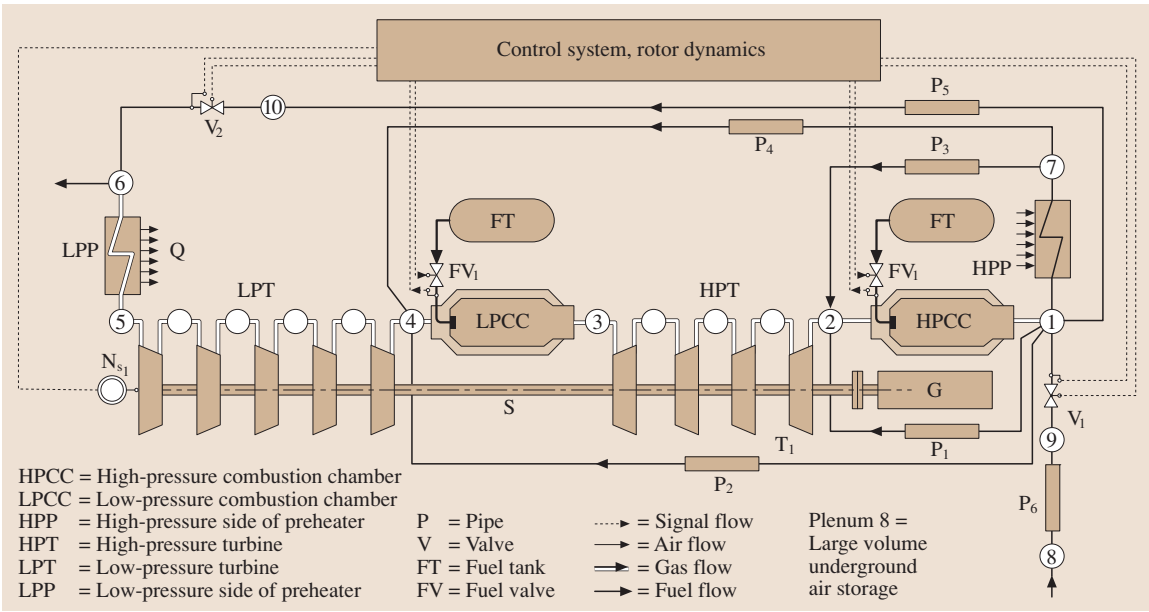


Fig. 12.35 Simulation schematic of the CAES shown in Fig. 12.34

shutdown was simulated, assuming a failure of the control system. This circumstance necessitates an intervention by the hydraulic emergency system. This incident simulates an extreme transient process within some of the components, as explained briefly. After the generator trip, the rotor is strongly accelerated because of the full turbine power acting on it (Fig. 12.36a).

The hydraulic emergency system intervenes only when the speed corresponding to the hydraulic emergency overspeed trip is reached. This intervention involves closing the fuel valves, FV_1 and FV_2, and air valves V_1, after which the system no longer receives any energy from outside (Fig. 12.36b). It also involves opening the bypass/blow-off valve V_2, which allows the high-pressure air contained in both large-volume combustion chambers as well as in the HP side of preheater to discharge.

The closing process of the inlet and shutdown valves and the opening of the bypass valves are shown in Fig. 12.36b. This process results in a steady drop in plena pressures and temperatures. As Fig. 12.37 shows, the pressure drop in the high-pressure section is initially

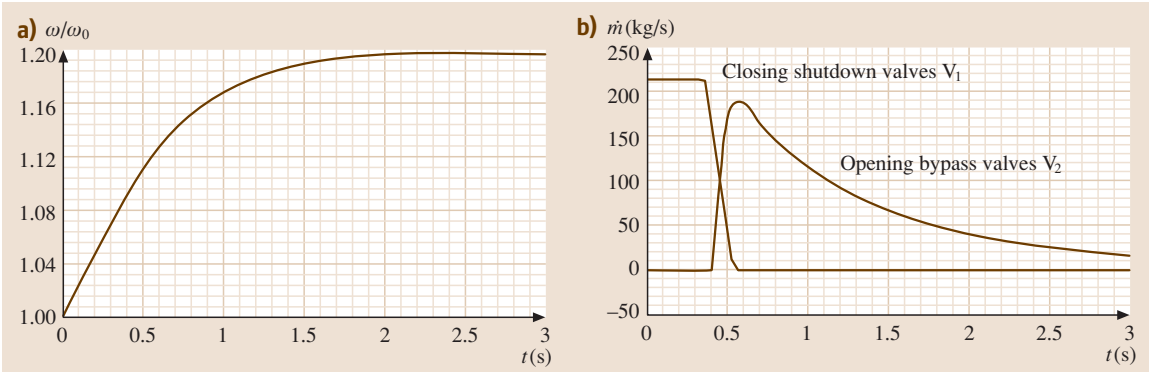


Fig. 12.36 **(a)** Relative angular velocity **(b)** and mass flows as functions of time. The inlet shutdown valves V_1 remain open until the trip speed at $t = 0.35$ s has been reached. The same procedure is true for opening the blow-off valves V_2. Closing the shutdown valves follow the ramp shown in **(b)**

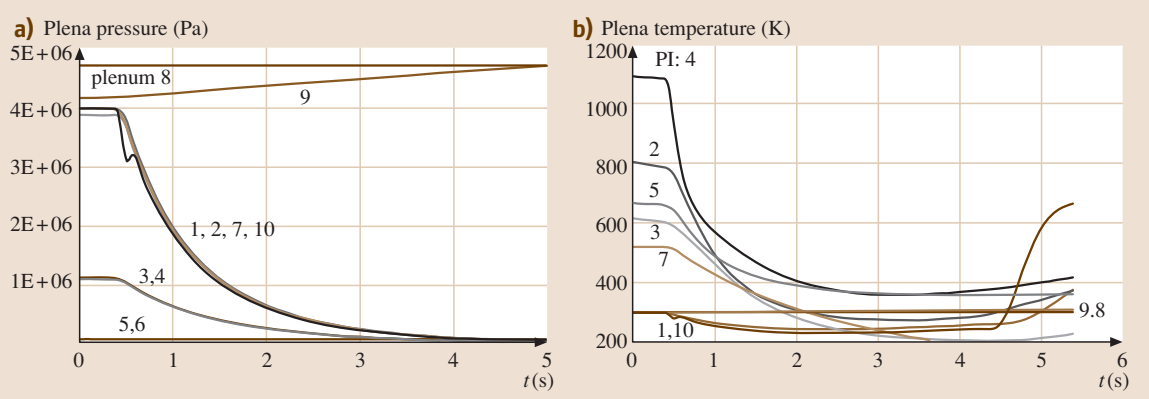


Fig. 12.37 (a) Plena pressure and (b) temperature as functions of time. The shutdown process causes rapid depressurization in the high-pressure plena 1, 2, 7, and 10

steeper than in the low-pressure section. This means that the enthalpy difference of the high-pressure turbine is reduced more rapidly than that of the low-pressure turbine. Immediately after the blow-off valve is opened, an abrupt pressure drop takes place in plenum 10, which is connected to plenum 1 via pipe P_5. Thereafter, dynamic pressure equalization takes place between the two plena. This drop in pressure and temperature causes a corresponding drop in the shaft power and the mass flow throughout the engine. Figure 12.38 shows the resulting drop in turbine inlet and exit temperature. The continuous decrease in turbine mass flow causes a strong dissipation of shaft power, resulting in the excessive increase of turbine exit temperature. In order to avoid thermal damages to the blades, a small stream of cold air is injected into the turbine flow path, which causes a reduction in temperature gradient. This is shown in Fig. 12.38 for the exit temperature at $t = 3.4$ s.

Dynamic behavior of the rotor speed is generally determined by the turbine power acting on the rotor. How the rotor behaves in response to a generator trip depends, in particular, on how long the full turbine power is available, a process monitored by the control and safety monitoring system. When the control system functions normally, a trip is signaled without delay to the shutdown valve. Failure of the control system causes the hydraulic emergency system to intervene. The intervention begins only when the speed corresponding to the hydraulic emergency overspeed trip is reached. During this process, and also the subsequent valve dead time, the rotor receives the full turbine power. The closing phase is characterized by a steady reduction in energy input from outside, which finally becomes zero. The total energy of the gases still contained in the system is converted by the two turbines into mechanical energy, causing the rotor speed to increase steadily (Fig. 12.36). When the instantaneous turbine power is just capable of balancing the friction and ventilation losses, the rotor speed reaches its maximum, after which it begins to decrease. Reducing the turbine mass flow (Fig. 12.39a) below the minimum value discussed in [12.6] causes the shaft power to dissipate completely as heat, resulting in negative values, as shown in Fig. 12.39b. From this point on, the rotational speed starts to decrease. Figure 12.39a depicts the mass flows through the HP and LT turbines that generate the total shaft power (Fig. 12.39b).

Case Study 2: Power Generation Gas Turbine Engine

The subject of this case study is the dynamic simulation of a BBC-GT9 gas turbine which is a single-spool single-shaft power generation gas turbine engine. It is utilized as a stand-alone power generator or in conjunction with combined cycle power generation. The engine shown in Fig. 12.40 consists mainly of three compressor stage groups, a combustion chamber, a turbine, a control system, and a generator.

The simulation schematic of this engine is presented in Fig. 12.41. The rotor speed and turbine inlet temperature are the input parameters for the controller, its output parameters are the fuel mass flow (fuel valve opening), and the mass flows of the bypass valves (bypass valve opening). The dynamic behavior of BBC-GT9 was experimentally determined for transient tests with extreme changes in its load. Its transient data was accurately documented by *Schobeiri* [12.9]. Starting from a given

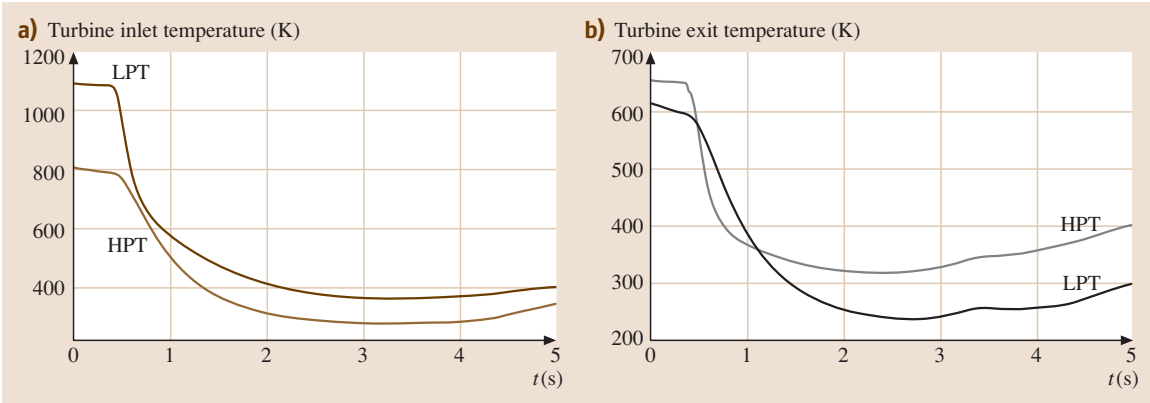


Fig. 12.38 (a) Turbine inlet and (b) exit temperature as functions of time. Note the changes of the exit temperature at $t = 3.4\,\mathrm{s}$

network load schedule, the dynamic behavior of the gas turbine is predicted and the results are presented.

The engine under consideration consists mainly of three compressor stage groups, a combustion chamber, a turbine, a control system, and a generator. The simulation schematic of this engine is similar to that of Fig. 12.42. For dynamic simulation, the first, second, and third stage groups are simulated using the row-by-row technique [12.6]. A similar row-by-row calculation procedure is applied to the turbine component. The rotor speed and the turbine inlet temperature are the input parameters for the controller; its output parameters are the fuel mass flow (fuel valve opening) and the mass flows of the bypass valves (bypass valve opening).

Simulation of an Adverse Dynamic Operation. Starting from steady state, in accordance with the load schedule indicated in Fig. 12.42a (curve 1), after 1 s, a generator loss of load is simulated that lasts for 6 s. The rotor at first reacts with a corresponding increase in rotational speed (Fig. 12.42b), which results in a rapid closing of the fuel valve (Fig. 12.42b). The rotational speed is then brought to an idling point and held approximately constant. The process of control intervention lasts until a constant idling speed is attained. After that, there is an addition of load in sudden increases, such that the gas turbine is supplying approximately 25% of its rated load (Fig. 12.42). The rotor first reacts to this addition of load with a sharp decrease in rotational speed, as exhibited in Fig. 12.42, causing a quick opening of the fuel valve (Fig. 12.42). After completion of the transient process, the steady off-design state is reached.

Plena Pressure and Temperature Transients. The above adverse dynamic operation triggers temporal changes of the flow quantities within individual com-

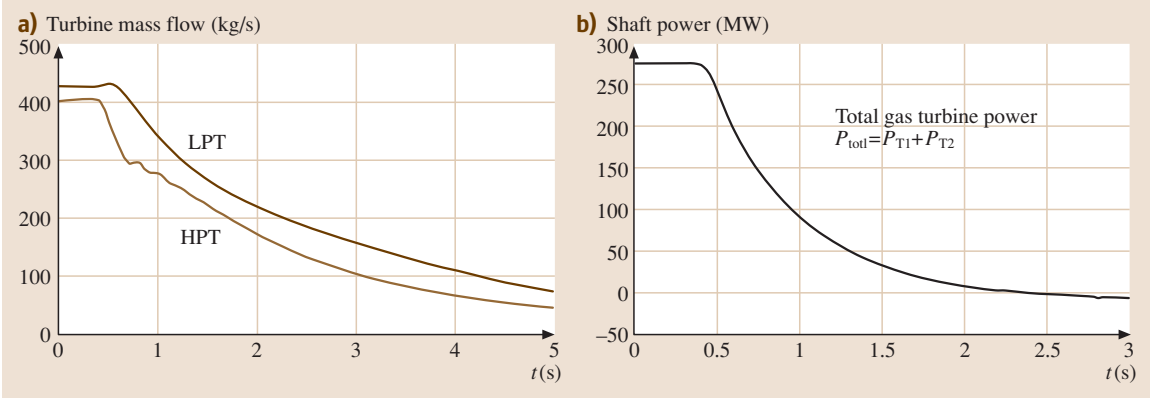


Fig. 12.39 (a) Turbine mass flow and (b) shaft power as functions of time

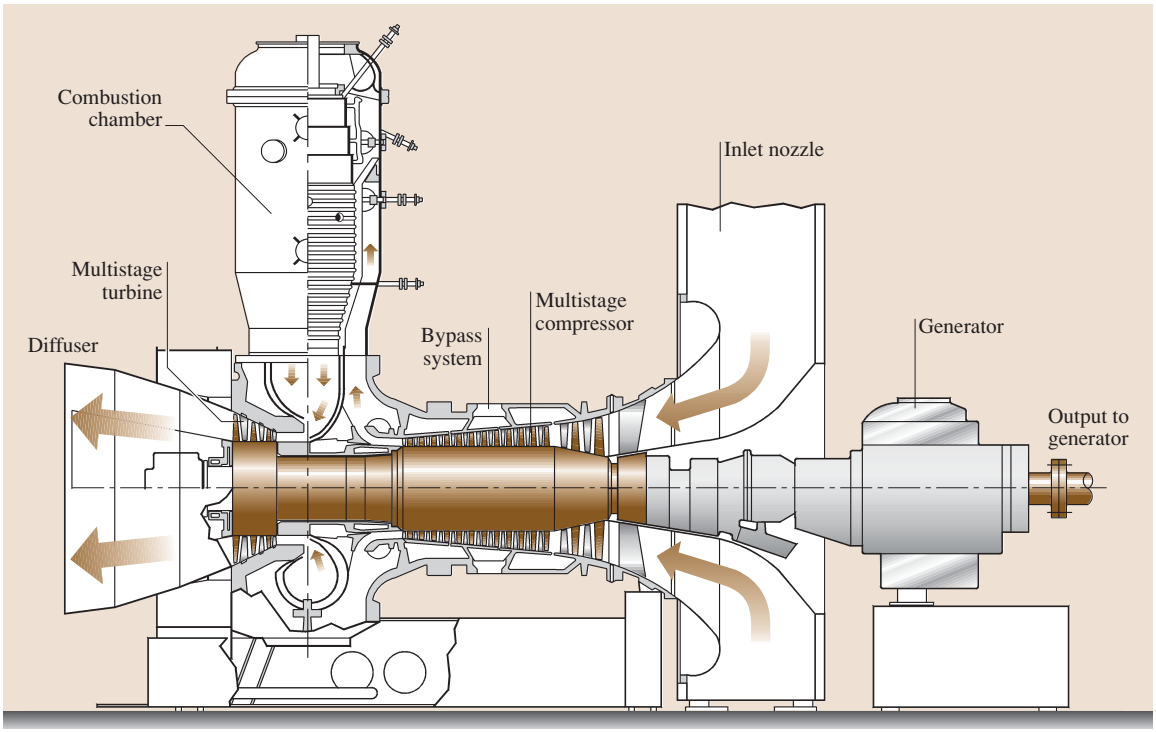


Fig. 12.40 A single-spool power generation gas turbine BBC-GT9

ponents. Figure 12.43 shows how the plena pressure and temperature change with time. Decrease of turbine power and increase of the shaft speed (Fig. 12.42) has caused the HP compressor exit pressure in plenum 5 to decrease. The temperature at the combustion chamber exit, plenum 6, and turbine exit, plenum 7, follow the course of fuel injection shown in Fig. 12.47b. The plena temperature upstream of the combustion chamber are not affected.

Compressor and Combustion Chamber Mass Flow Transients. Figure 12.44 exhibits the mass flow transients through low pressure (LP), intermediate pressure (IP), and high pressure (HP) compressors. While the IP- and HP-stage groups have the same mass flow, the LP part has a greater mass flow. The difference of 1 kg/s is due to the cooling mass flow extraction. As briefly mentioned, the increase in shaft speed and the simultaneous decrease in compressor power consumption leading to

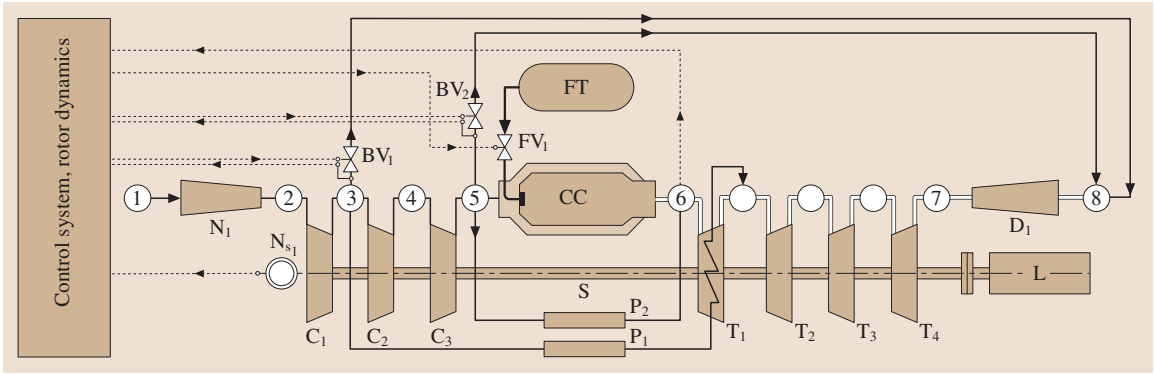


Fig. 12.41 Simulation schematic of BBC-GT9 shown in Fig. 12.40

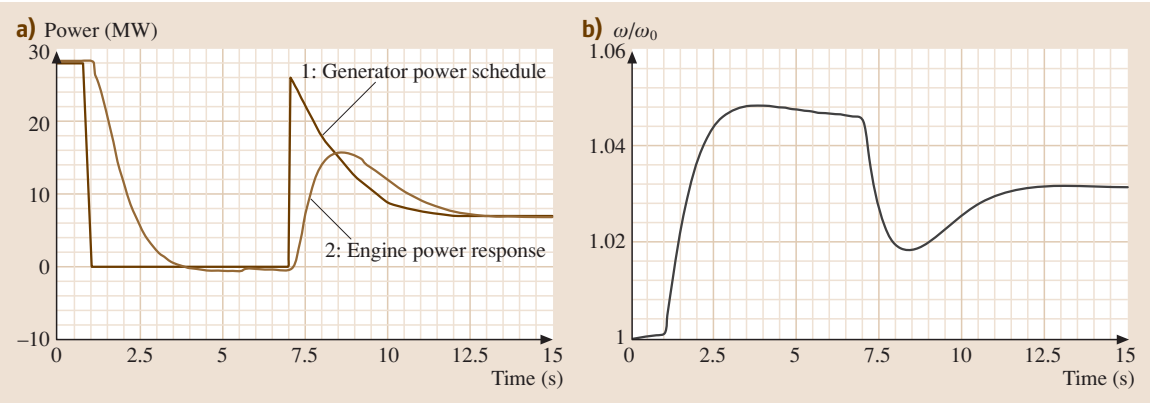


Fig. 12.42 **(a)** Generator load schedule (*curve* 1). Sequence of events: steady operation from 0 to 1 s, sudden loss of load, idle operation, sudden addition of load, continuous decrease of load to 25%. *Curve* 2: engine power response. **(b)** Relative shaft speed as a function of time

the compressor pressure drop has caused an increase in the compressor mass flow during the process of loss of load that lasts up to $t = 6$ s. The sudden load addition reduces the compressor mass flow.

The combustion chamber mass flow shows a similar course with a substantial difference: a substantial portion of the compressor mass flow is extracted for combustion chamber exit temperature mixing cooling.

Combustion Chamber Gas and Metal Temperature Transients. The combustion chamber component used in this simulation has three segments that separate the primary combustion zone from the secondary cooling air zone. Its module is shown in Fig. 12.45. Figure 12.46 exhibits the combustion chamber gas and metal temperatures as functions of time. Compressed air enters the combustion chamber at station 1 (Fig. 12.46). Fuel is added and the segment cooling occurs according to the procedure described in [12.6]. The secondary mass flow portions $\dot{m}_{Si}$ serve as cooling jets and are mixed with the combustion gas, thus reducing the gas temperature. Before exiting, the combustion gas is mixed with the mixing air stream $\dot{m}_M$, further reducing the temperature. Figure 12.46b shows the mean segment temperatures. In accordance with the measurements on this gas turbine, the flame length extends from station 1 to 3, which makes segment number 2 the hottest one. We assumed that all secondary cooling channels are open.

Turbine and Fuel Mass Flow Transients. Figure 12.47a exhibits the turbine mass flow transient, which is dictated by the compressor dynamic operation. The difference between the turbine and the compressor mass flow is the injected fuel mass flow. The particular course

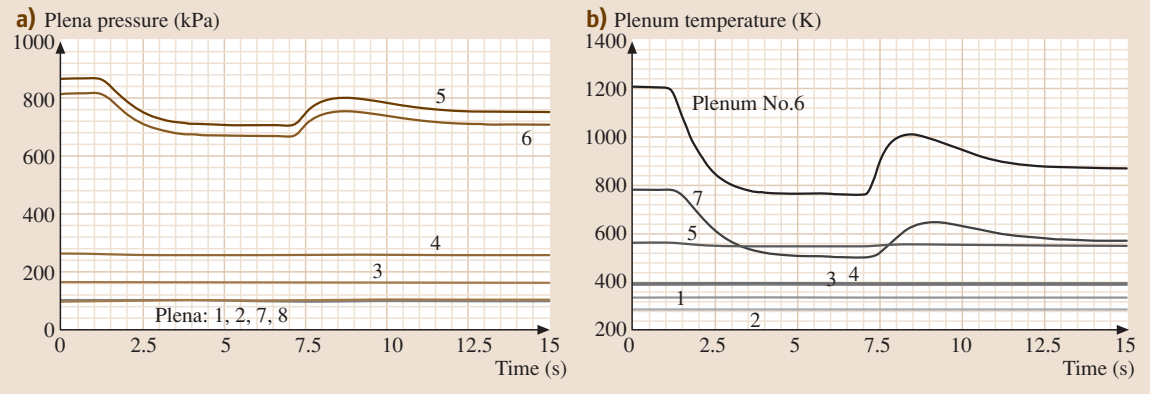


Fig. 12.43 **(a)** Plena pressure and **(b)** temperature as functions of time. Individual plena are labeled

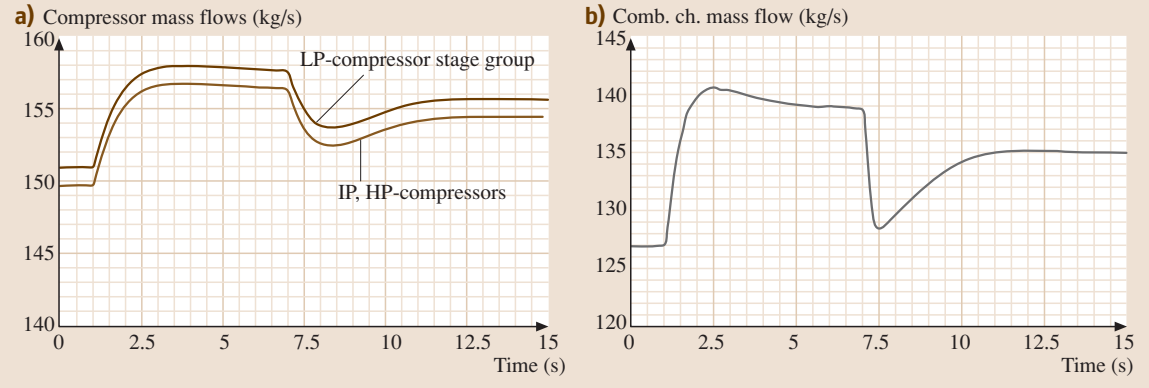


Fig. 12.44 (a) Compressor and (b) combustion chamber mass flows as functions of time

of fuel mass flow shown in Fig. 12.47b is due to the intervention of the control system. An increase in rotational speed causes the controller to close the fuel valve. Subsequent addition of generator load results in a steep drop of rotational speed, which causes an opening of the fuel valve.

Case Study 3: Simulation of a Multispool Gas Turbine Engine

The subject of this study is the nonlinear dynamic simulation of a gas turbine engine with a higher degree of complexity than the previous cases. For this purpose a three-spool thrust-generating gas turbine engine is designed that incorporates advanced components. The three-spool four-shaft high-performance gas turbine engine consists of a low-pressure spool that incorporates the LP compressor and turbine connected via shaft S_1. The intermediate pressure spool integrates the IP compressor and turbine connected via shaft S_2. The high-pressure spool carries the HP compressor and HP turbine on shaft S_3. To increase the level of engine complexity, a fourth shaft, S_4, with the power generating turbine T_4, was attached to the exit of the three-spool gas generating unit as shown in Fig. 12.48. The transient operation is controlled by a given fuel schedule. The component nomenclature for this configuration is the same as for the previous cases. The simulation schematics shown in Fig. 12.48 represents the modular configuration of the gas turbine.

Fuel Schedule and Rotor Response. The dynamic behavior of the above engine is simulated for an adverse acceleration–deceleration procedure (Fig. 12.49).

The transient operation is controlled by an open-loop fuel schedule shown in Fig. 12.49a. The three spools and the fourth shaft run independently at different rotational speed (Fig. 12.49b). The fuel schedule generated completely arbitrarily simulates an acceleration–deceleration procedure with emphasis on deceleration. We start with the steady-state operation and reduce the fuel mass flow to $\dot{m}_F = 2.8\,\mathrm{kg/s}$ for about 2 s. During this short period of time, the engine operates in a dynamic state which is followed by a cyclic acceleration–deceleration event with the ramps given in Fig. 12.49. The dynamic operation triggers a sequence of transient events within individual components that are discussed in the following sections.

Rotor Speed Behavior. The transient behavior of the three spools as well as the power shaft is determined by the net power acting on the corresponding rotor. For each individual spool, the cyclic acceleration–deceleration event has caused a dynamic mismatch between the required compressor power consump-

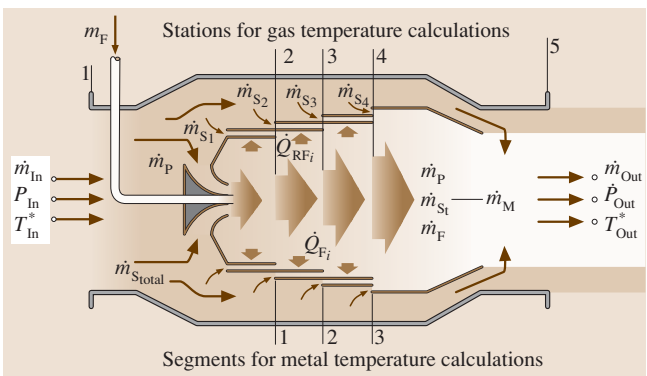


Fig. 12.45 Combustion chamber module, stations and segments, $\dot{m}_P$ – primary $\dot{m}_{S_{tot}}$ – total secondary air, $\dot{m}_{Si}$ – individual secondary air

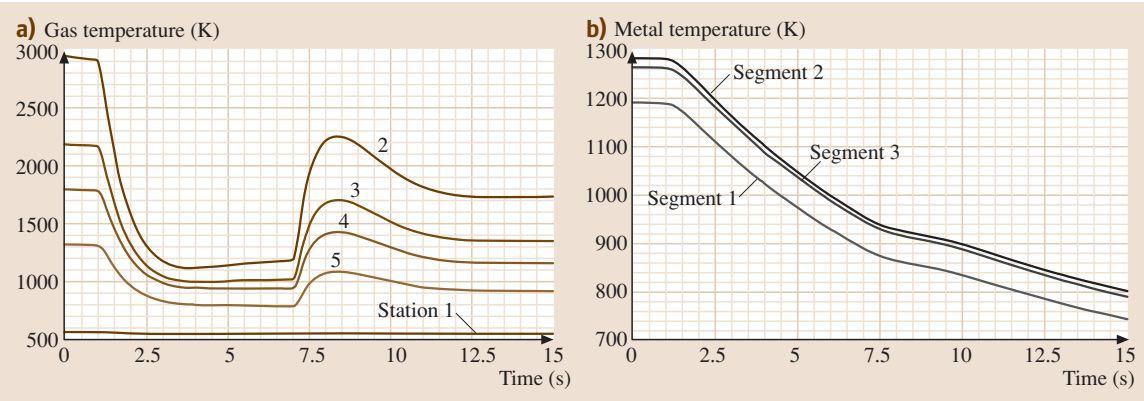


Fig. 12.46 (a) Combustion chamber gas and (b) metal temperature at different positions as functions of time

tion and the turbine power generation, as shown in Fig. 12.50. While the LP and IP spools 2 and 3 decelerate under the influence of negative net power, the HP spool 3 reacts faster to the acceleration. Since the fuel schedule places special weight upon deceleration, the rotational speeds of all three spools have a decelerating tendency as shown in Fig. 12.50.

Pressure and Temperature Transients within Plena. The change in fuel mass flow triggers a chain of transient events within the plena, as shown in Fig. 12.51. Plena pressure 5 and 6, which corresponds to the exit pressure of the HP compressor and the combustion chamber, are strongly affected by the cyclic fuel change, whereas the other plena that correspond to the inlet and exit plena of the remaining components experience moderate changes. The plena temperature distributions downstream of the combustion chamber shown in Fig. 12.51b reflect the course of the fuel schedule.

Combustion Chamber Gas and Metal Temperature Transients. Figure 12.52 exhibits the combustion chamber gas and metal temperatures as functions of time. The combustion chamber component used in this simulation has three segments that separate the primary combustion zone from the secondary cooling air zone. Its module is shown in Fig. 12.47. Compressed air enters the combustion chamber at station 1 (Fig. 12.52a). Fuel is added and the segment's cooling occurs according to the procedure described in [12.6]. The secondary mass flow portions serve as cooling jets and are mixed with the combustion gas, thus reducing the gas temperature. Before exiting, the combustion gas is mixed with the mixing airstream, further reducing the temperature. Figure 12.52a shows the mean segment temperatures.

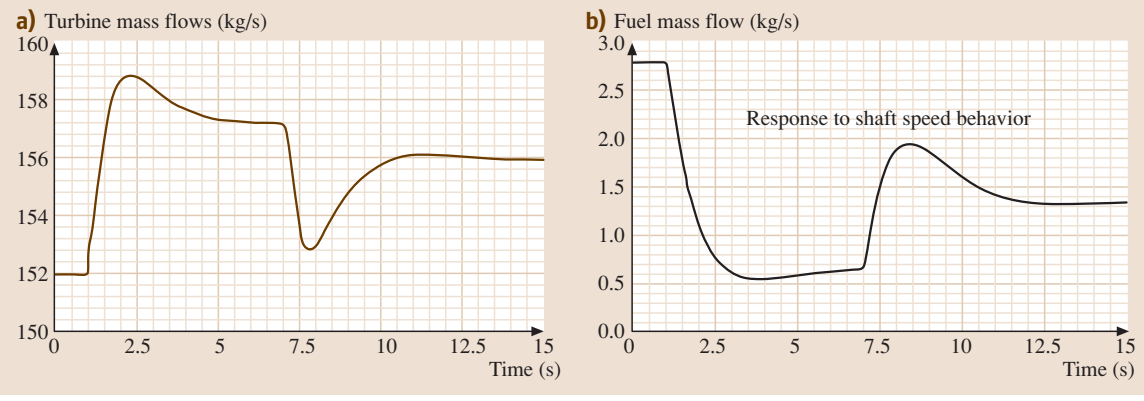


Fig. 12.47 (a) Turbine and (b) fuel mass flow as functions of time. The fuel mass flow is controlled by the shaft rotational speed

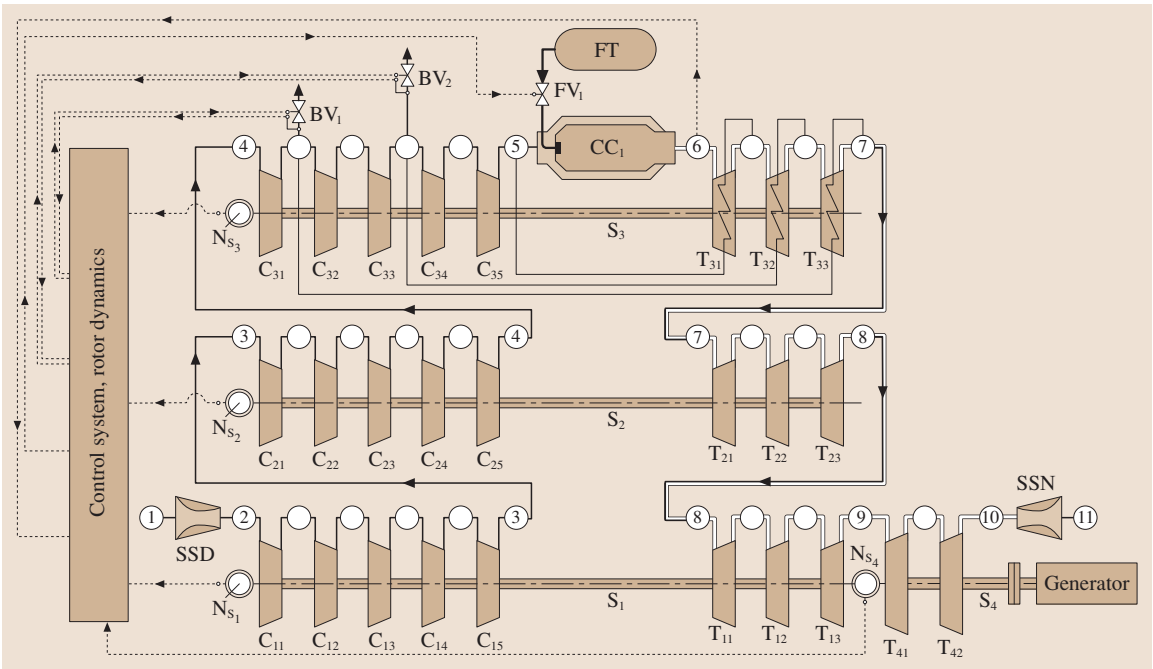


Fig. 12.48 Simulation schematic of three-spool four-shaft high-performance gas turbine engine. Spool 1 incorporates the LP compressor and LP turbine connected via shaft S_1; spool 2 incorporates the IP compressor and IP turbine connected via shaft S_2; spool 3 incorporates the HP compressor and HP turbine connected via shaft S_3

The flame length extends from station 1 to 3, which makes segment number 2 the hottest one. As seen, the gas temperature at station 2 follows the sharp changes in the fuel schedule. By convecting downstream, these sharp changes are smoothed out. The wall temperatures shown in Fig. 12.52b exhibit similar tendencies.

Compressor and Turbine Mass Flow Transients. Figure 12.53a exhibits the compressor mass flow transients, which are dictated by the compressor dynamic operation. The difference in compressor mass flow is due to the mass flow extraction for cooling purposes. Turbine mass flows are illustrated in Fig. 12.53b. Except for a minor time lag, they show identical distributions. The

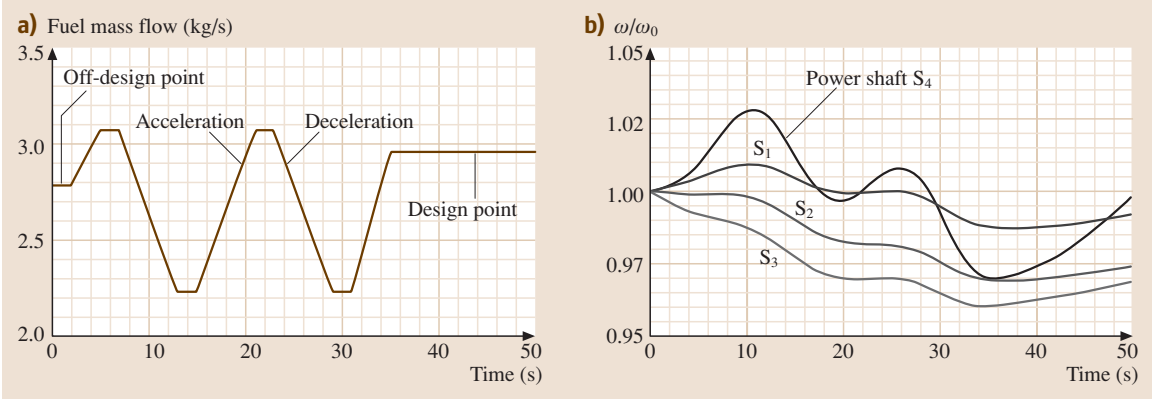


Fig. 12.49 **(a)** Fuel schedule starts with an off-design mass flow followed by a cyclic acceleration–deceleration procedure. **(b)** Rotational speed of the three spools and the power shaft

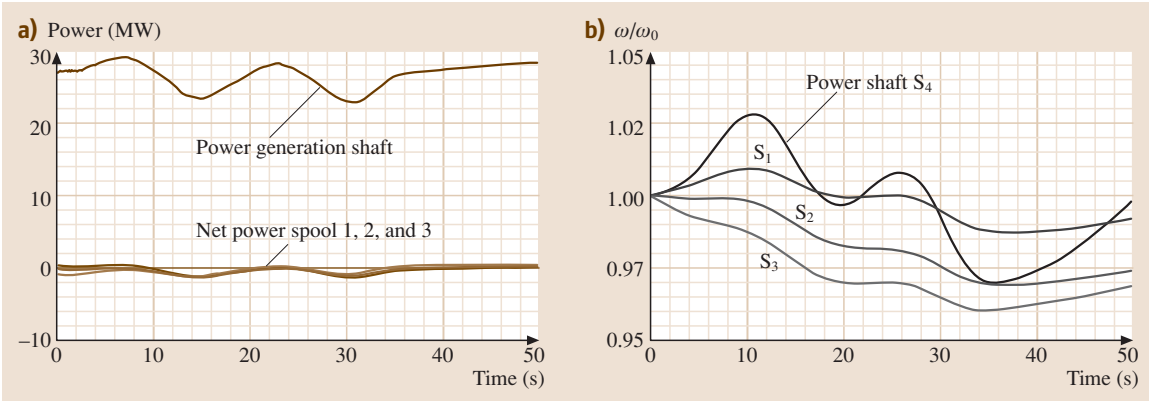


Fig. 12.50 (a) Net power acting on the three spools causing a dynamic mismatch and the power generated by the fourth shaft. (b) Relative rotor speed of three spools and the fourth shaft

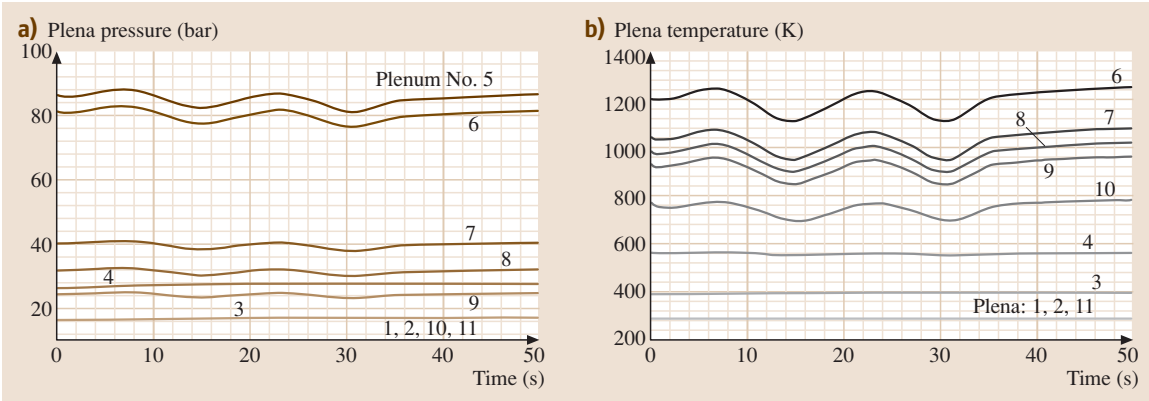


Fig. 12.51 (a) Plena pressure and (b) temperature as functions of time

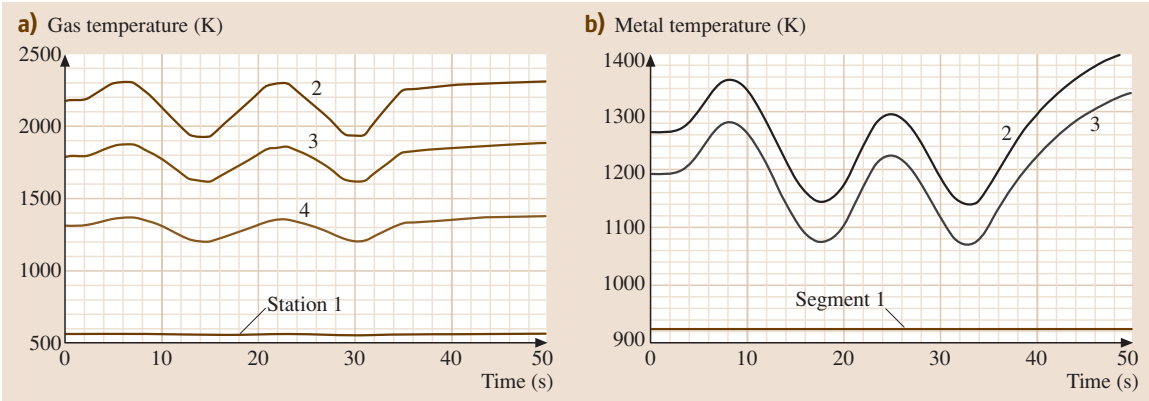


Fig. 12.52 (a) Combustion chamber gas and (b) metal temperature as functions of time

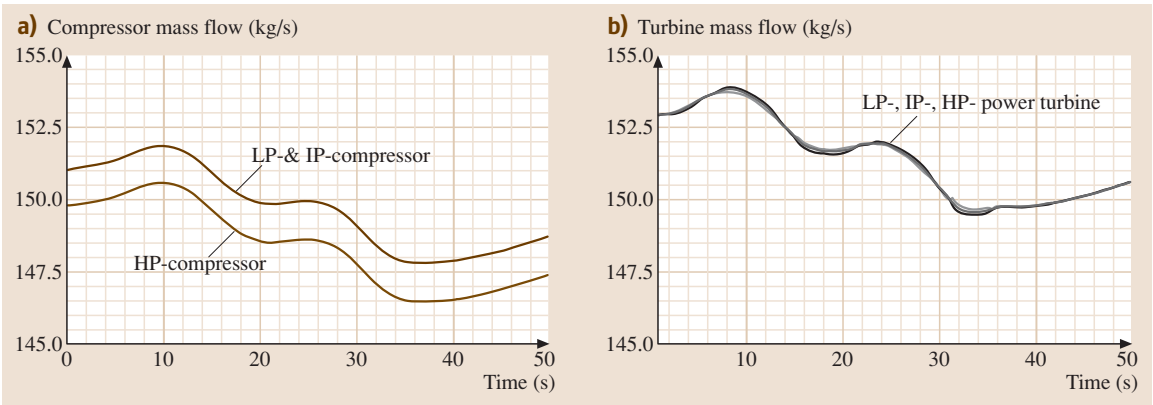


Fig. 12.53 **(a)** Compressor and **(b)** turbine mass flow as functions of time

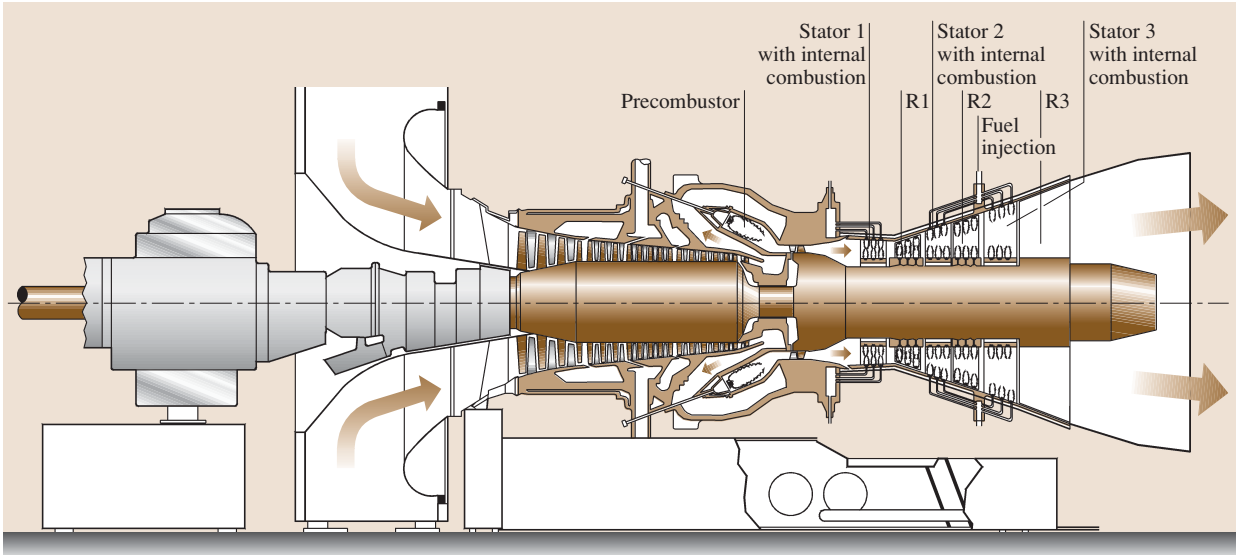


Fig. 12.54 A derivative of the ultrahigh-efficiency gas turbine engine with a multistage compressor, a conventional combustion chamber PC, a single-stage reheat turbine RT, a three-stage turbine with an integrated stator internal combustion. B1, B2: compressor bypass blow-off, Fl1, Fl2: fuel lines to stator. The combustion process takes place inside the precombustor and stator flow path (after [12.7])

difference in turbine and compressor mass flow is due to the addition of fuel.

12.2.6 New Generation Gas Turbines, Detailed Efficiency Calculation

One of the interesting aspects of a dynamic simulation is the capability to calculate the gas turbine thermal efficiency dynamically during steady-state and dynamic operation. Such calculations are performed to compare the thermal efficiencies of four gas turbines with different design methodologies. The calculations are performed with the nonlinear dynamic code GETRAN©. The first gas turbine dynamically simulated for efficiency calculation is a conventional single-shaft, single-combustion chamber power generation gas turbine. The second one is the ABB GT 24/26. The third is an ultrahigh-efficiency gas turbine (UHEGT) with a precombustor, a reheat turbine stage, and an integrated stator internal combustion, as

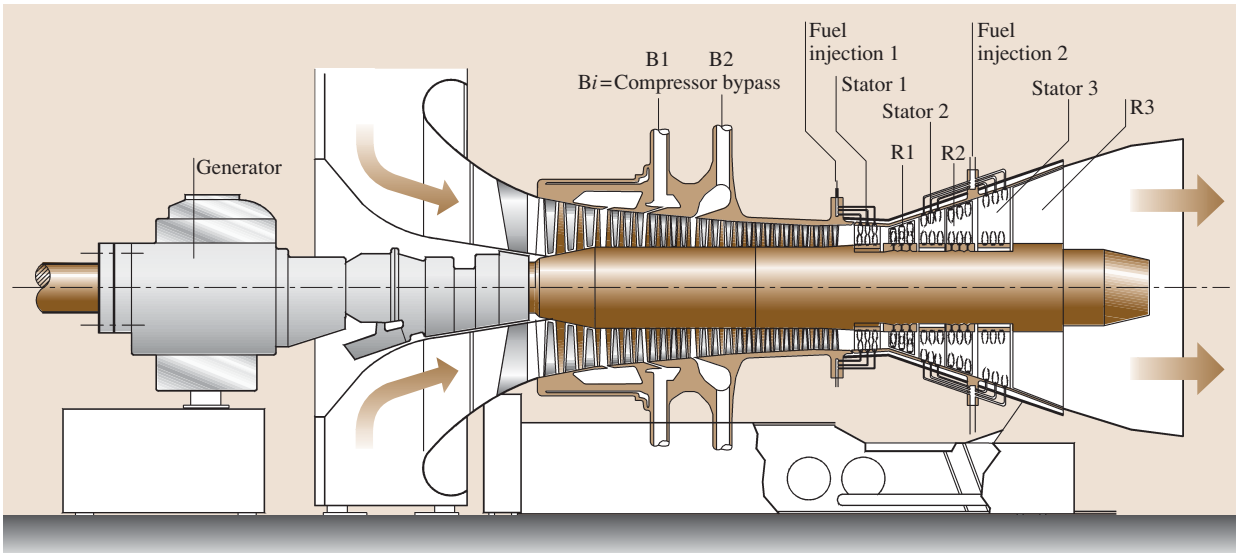


Fig. 12.55 An ultrahigh-efficiency gas turbine engine with a multistage compressor, and a three-stage turbine with an integrated stator internal combustion. B1, B2: compressor bypass blow-off, Fl1, Fl2: fuel lines to stator. The combustion process takes place inside the stator flow path (after *Schobeiri* [12.7])

illustrated in Fig. 12.54. For the fourth gas turbine, the combustion process is placed entirely within the stator rows, thus eliminating the combustion chambers all together, Fig. 12.55. The dynamic efficiency calculation results are presented in Fig. 12.56. To accurately determine the thermal efficiency and specific work of the gas turbines, calculations are performed with GETRAN© and the results are presented in Fig. 12.56. To compare the degree of efficiency improvement, the thermal efficiency and specific work of a baseline GT, GT-24, and the three UHEGT gas turbines are included in the figures.

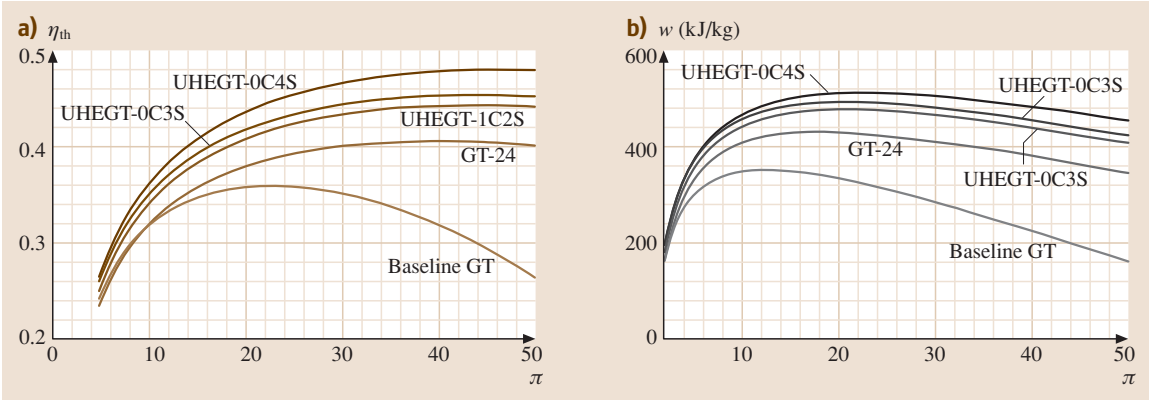


Fig. 12.56 (a) Thermal efficiency and (b) specific work as functions of compressor pressure ratio for the reference baseline gas turbine engine, the ABB-GT-24 with two combustion chambers, the UHEGT-1C2S with one conventional combustion chamber and two UHEGT-stator combustion, the UHEGT-03S with three-stator combustion, and the UHEGT-0C4S with four-stator combustion. Turbine inlet temperature = 1200 °C, calculation with GETRAN (after [12.9])

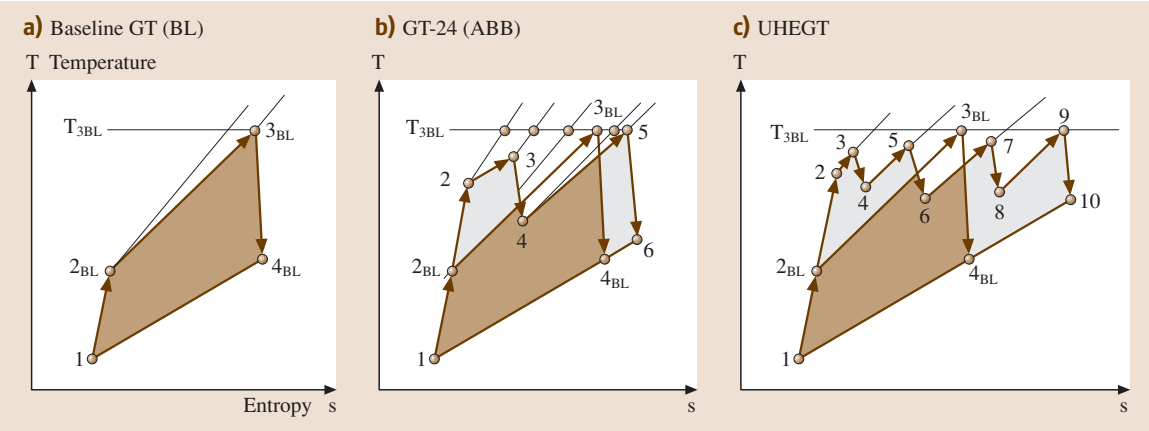


Fig. 12.57a–c Process comparison for (a) the baseline GT, (b) the GT-24, (c) the ultra high efficiency gas turbine technology (UHEGT) which has four stages with four integrated stator internal combustion

For a UHEGT with three-stator combustion, denoted by the curve UHEGT-0C3S, a thermal efficiency above 45% is calculated. This exhibits an increase of at least 5% above the gas turbine engine GT-24, which is close to 40.5%, as shown in Fig. 12.56a. Increasing the number of stators for internal combustion to four (curve labeled "UHEGT-0C4S") raises the efficiency above 48%. This is an enormous increase compared with any existing gas turbine engine. In the course of this calculation, the UHEGT technology is applied to a gas turbine engine with a precombustion chamber, such as the first one in GT-24 (Fig. 12.54). Using this combustion chamber with two-stator combustion, the curve labeled "UHEGT-1C2S" shows an efficiency of 44%. This is particularly interesting for upgrading the existing gas turbines with UHEGT technology. Figure 12.56b shows a comparison of the specific work for the gas turbines discussed above. Compared to the GT-24, UHEGT technology has about 20% higher specific work, making these engines very suitable for aircraft, stand-alone, as well as for combined cycle power generation applications. The evolution of the gas turbine efficiency improvement is summarized in the corresponding T–s-diagrams presented in Fig. 12.57a–c. In addition to the efficiency improvement brought about by utilizing a reheat turbine and two combustion chambers already discussed in conjunction with Fig. 12.26a,b further improvement is achieved by using stator internal combustion. As shown in Fig. 12.54c, with the UHEGT facilitates *quasi-Carnotising the Brayton cycle*. It is interesting to note that this efficiency increase can be established at a compressor pressure ratio of $\Pi_{\text{UHEGT}} \approx 35–40$, which can be achieved easily by existing compressor design technology with high polytropic efficiency. In performing the GETRAN© calculation, compressor and turbine efficiencies are calculated on a row-by-row basis. This automatically accounts for an increase of secondary flow losses based on aspect ratio decrease. Thus, in a compressor case, efficiency decrease with pressure ratio increase is inherently accounted for.

References

12.1 M.T. Schobeiri: *Turbomachinery Flow Physics and Dynamic Performance* (Springer, Berlin, Heidelberg 2004)

12.2 M.H. Vavra: *Aero-Thermodynamics and Flow in Turbomachines* (Wiley, New York 1960)

12.3 W. Traupel: *Thermische Turbomaschinen*, Vol.1 (Springer, Berlin, Heidelberg 1977)

12.4 J.H. Horlock: *Axial Flow Compressors* (Butterworth, London 1966)

12.5 J.H. Horlock: *Axial Flow Turbine* (Butterworth, London 1966)

12.6 M.T. Schobeiri: Digital computer simulation of the dynamic operating behavior of gas turbines, J. Brown Boveri Rev. **74**(3) (1987)

12.7 M.T. Schobeiri: The ultra-high efficiency gas turbine engine with stator internal combustion, UHEGT Patent 1389-TEES-99 (1999)
12.8 M.T. Schobeiri, S. Attia: Advances in nonlinear dynamic engine simulation technology, ASME 96-GT-392, Int. Gas Turbine Aero-Engine Congress Exposition (Birmingham 1996)
12.9 M.T. Schobeiri, M. Abouelkheir, C. Lippke: GETRAN: A generic, modularly structured computer code for simulation of dynamic behavior of aero- and power generation gas turbine engines, ASME Trans. J. Gas Turbine Power **1**, 483–494 (1994)

13. Transport Systems

Gritt Ahrens, Torsten Dellmann, Stefan Gies, Markus Hecht, Hamid Hefazi, Rolf Henke, Stefan Pischinger, Roger Schaufele, Oliver Tegel

Transportation is derived from two Latin words *trans* and *porta* meaning *in between* and *carrying*, respectively. Transportation is seen as one of the basic human needs and has a significant impact on a country's economy; productivity usually correlates well with the amount of transportation of goods and people. Transportation takes place on the ground, sea, and in the air and can be subdivided into the areas automotive, railway, naval, and aerospace. The respective engineering disciplines have gained increasing importance in the past as they face severe challenges for the future arising from: (i) increased transportation demand and customer needs, (ii) shortage of energy and rising fuel prices, and (iii) more stringent legislative requirements regarding for example pollutant and noise emissions and safety issues.

Section 13.2 provides an overview of aspects of automotive engineering. It starts with a historical view of how cars have evolved over time until today. Section 13.2.2 covers automotive technology, first describing the different car types and the fundamental requirements for car development. The technological areas and corresponding components relevant to cars are then briefly explained according to their major functions, the requirements they have to fulfill, and the challenges for further development in the future.

The car development process, with emphasis on the early phase where the car concept is defined and verified, is described in Sect. 13.2.3. Finally, some methods used in car development and cross-functional aspects to be covered in order to manage the car development process and meet the goals of a car development project are depicted.

At the end of the Chapter, a list of references is provided which will enable the interested reader to obtain detailed information about the technological aspects of modern cars and their development.

For general reading on automotive engineering, refer also to [13.1–5].

13.1 Overview

The forecasted increase in the world's population and its industrialization are the main drivers for the predicted rise in transportation demand, posing environmental concerns but also possible energy supply risks for the future.

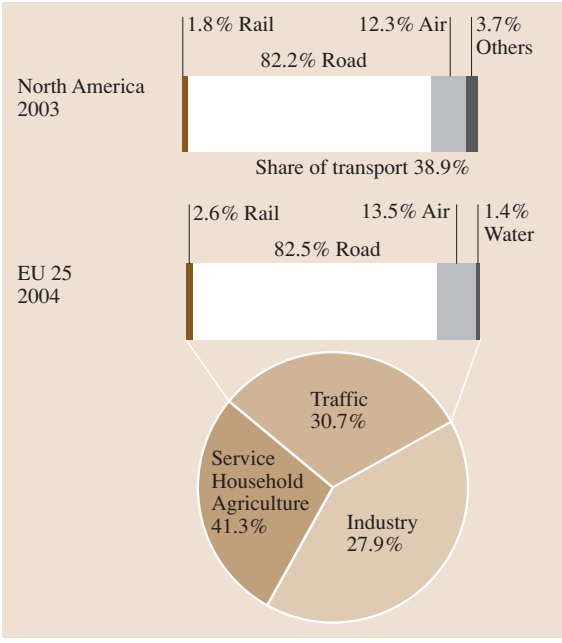


Fig. 13.1 Share of energy use in transport and distribution by modes of transport (from [13.6, 7])

Referring to Fig. 13.1, traffic's share in total demand for primary energy in the year 2004 amounted to 30.7% in Europe and to 38.9% in 2003 in North America (USA and Canada). Road transportation consumes the highest portion: 82.5% in Europe, and 82.3% in North America.

In Fig. 13.2 the predicted development in passenger travel and freight transportation in Europe up to the year 2020 is shown. For passenger travel, one can see a steady increase in road and air transport, while rail and public transportation remain more or less constant.

An increase in freight transportation is also foreseen, whereas the highest total share, in ton kilometers, is and will continue to be covered by road transportation, accounting for most of the increase up to the year 2020. However, it has to be noted that air transportation is forecasted to double in the same period.

In Fig. 13.3 the resulting predicted rise in energy demand of the transportation sector is shown for Organisation for Economic Co-operation and Development (OECD) and non-OECD countries.

For OECD countries an increase of 48% and for non-OECD countries a rise of 66% is envisioned up to

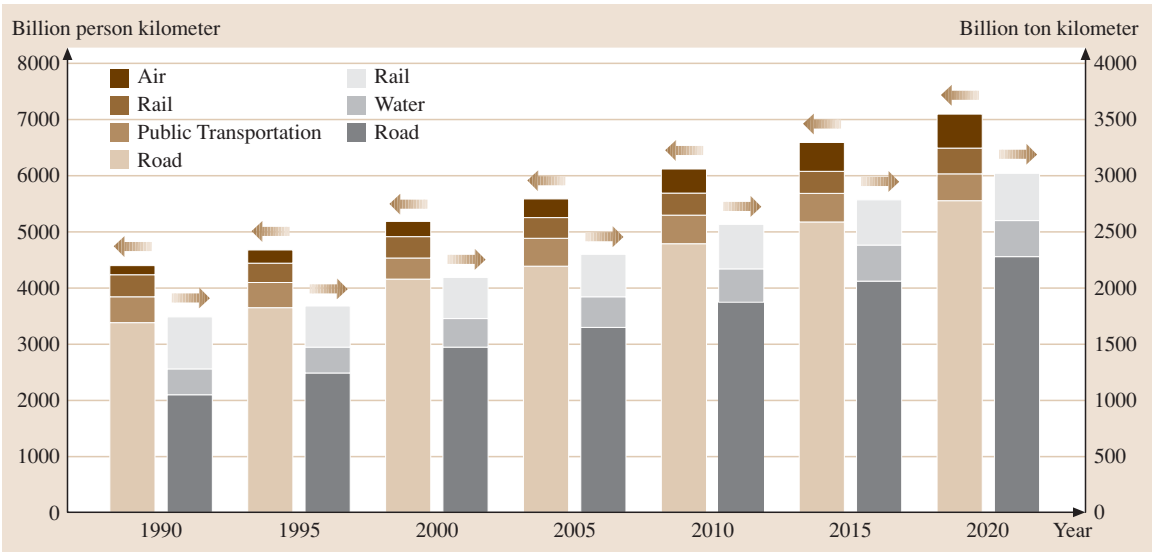


Fig. 13.2 Share of different modes of passenger and freight transportation in Europe (EU-25) (Source: European Commission)

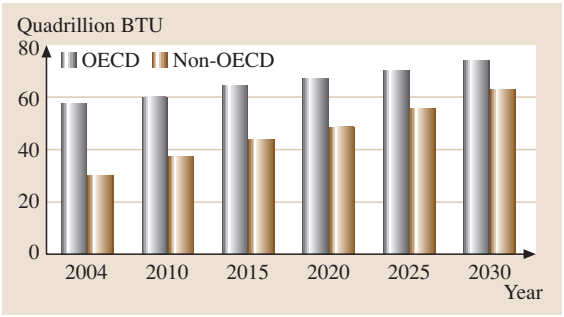


Fig. 13.3 OECD and non-OECD transportation, delivered energy consumption 2004–2030 (from [13.8])

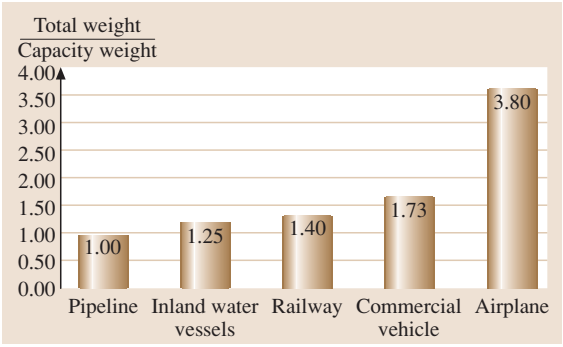


Fig. 13.5 Fraction of total weight to capacity weight for different transportation systems

the year 2030, leading to the necessity to improve efficiency of transportation as well as to search for new sources of energy.

An accurate comparison of the different transport modes (road transportation, aviation, and rail) regarding efficiency is difficult, because calculations are based on assumption for, e.g., travel distances, transport speeds, or average occupation.

In comparison with road transportation, rail transportation is more efficient. In [13.9], an electrical energy demand of 37 W h/seat-km is stated for an InterCity Express (ICE) (electro) with maximum speed higher than 200 km/h. Assuming an average passenger per seat ratio of 47.6%, an energy consumption of 78 W h/passenger-km results, which is equal to the gasoline energy content of 0.9 l/100 km. This energy does not include the efficiency of the energy production. Regarding aviation, in [13.10] the average kerosene consumption per 100 passenger-km of the Lufthansa fleet for intercontinental, continental, and regional flights are given as 4.06 l/100 passenger-km, 4.39 l/100 passenger-km, 8.85 l/100 passenger-km.

In Figs. 13.4 and 13.5 further criteria regarding transportation space and mass requirements are compared. Commercial vehicles and railways have higher transportation velocities compared with inland vessels and pipelines and can thus handle larger transportation streams. In contrast, the pipeline has the smallest profile area due to its construction and therefore achieves the largest transportation stream relative to its cross-sectional area.

The most favorable payload ratio is achieved with the pipeline, followed by inland vessels, railroads, commercial vehicles, and finally by airplanes.

Economical aspects such as energy and cost efficiency, investment needs, transportation speed, and space requirements will jointly play a decisive role for

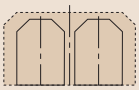
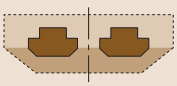
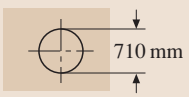
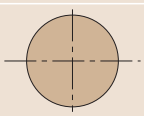
Means of transportation	Cross-sectional profile	Transportation speed v (mph)	Transportation flow $\dot{v}$ (m^3/h)	Profile area A_p (m^2)	Relative transportation flow $\nu = \dot{v}/A_p$
Railway		32 (without shunting operation)	20 000	37	
Highway		32	14 500	115	
Canal		7.5	6250	470	
Pipeline		4.5	2850	0.4	

Fig. 13.4 Comparison of freight transportation systems (from Institut für Kraftfahrzeuge (ika), Aachen)

which transportation mode will become dominant in which area.

In addition, legislative requirements pose increasing constraints on all areas of transportation. One such area is the restriction of pollutant emissions of vehicles such as nitrogen oxides, unburned hydrocarbons, carbon monoxide, and particulates, resulting from the combustion of hydrocarbon bases fuels with air. In Fig. 13.6 particulate (PM) and nitrogen oxides (NO_x) emissions limits of diesel-powered passenger vehicles are shown for the USA and Japan as well as for the European Union (EU). These graphs illustrate a common trend among the industrialized nations to reduce the limits of pollutants emitted by vehicles.

The current focus on reducing the climate-relevant CO_2 emissions, which are not regarded as toxic emissions, will most likely lead to legal limitations in the future as well.

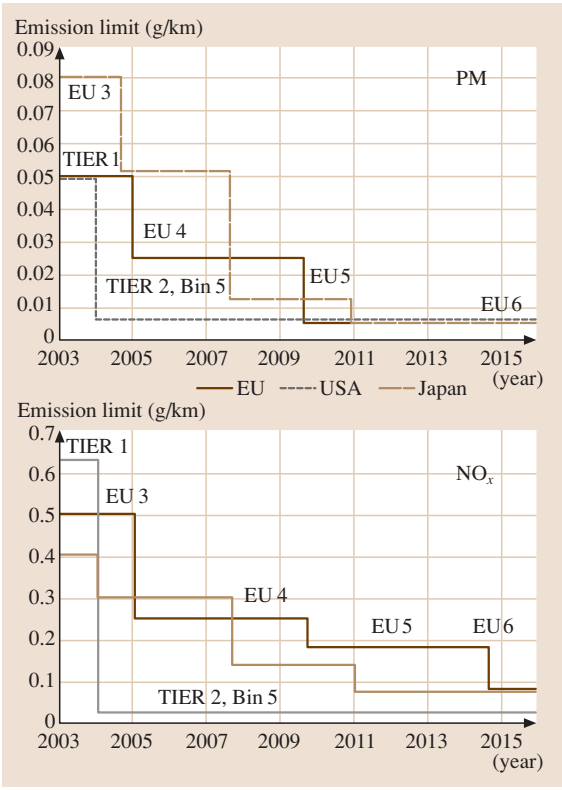


Fig. 13.6 Emission limits of diesel passenger cars (TIER 1 and TIER 2, bin 5: 50 000 miles; EU 5: draft proposal by European Commission in July 2005; Japan beyond model year 2011: proposed limits)

Besides emission of chemical components, noise is also generated and emitted by the various transport modes. Noise levels of 130–140 dB(A) are registered as pain; lower noise levels are also considered as harmful today, causing annoyance and aggression, high stress levels, hearing loss, and other harmful effects depending on the noise level and duration of exposure. Major sources of transport-related noise emissions are road, rail traffic, and aircraft traffic.

As early as 1937, noise limits for motor vehicles were specified in the German road traffic act (StVZO). The § 49 StVZO states: *Motor vehicles . . . must be created in such a manner that the noise produced does not exceed a value unavoidable by the current state of technology*. Over time, the limits have been tightened. A new EU regulation is in progress.

In 1971, the International Civil Aviation Organization (ICAO), an agency of the United Nations that codifies the principles and techniques of international air navigation, established regulations for limiting the noise radiation for civil aircrafts, distinguishing between different aircraft classes. As for road transportation, regulations for air traffic have also been tightened in several steps.

Finally, safety issues should be mentioned as a major driver for transportation technology, being imposed by legislation as well as being driven by customer requirements. In general, one has to differentiate between measures enabling a safe use of the transport mode involved and the possibility of harming a third party using the same, a different or no transport mode. For aeroplanes, for example, safe use is the main focus,

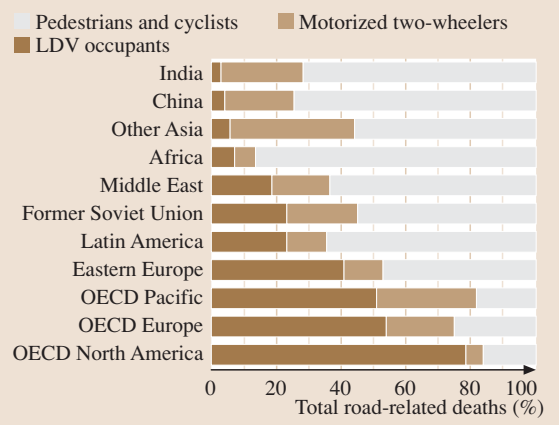


Fig. 13.7 Share of total road-related deaths by category of road users (source: 2007-09-14_WBCSD.org_mobility-full.pdf/LDV Light Duty Vehicle)

accomplished by stringent service routines and scrutinized development processes. On the one hand this is because any severe failure in air would almost certainly cause a large number of casualties and on the other hand design alterations of the aircraft cannot prevent harm to third parties. These have to be safeguarded within the operation, e.g., air-traffic control is responsible for ensuring that two aircraft do not collide. In contrary, road vehicles operate in partially very dense traffic and, although various regulations concerning servicing apply, they are mostly dependent on the awareness of the driver. In Fig. 13.7 the distribution of road-related deaths is shown. In North America light duty vehicles (LDV) contribute approximately 80% of road-related deaths, which in turn means that improving the vehicle's passive safety systems could be beneficial. In contrast to this, 70% of road-related deaths in India are accounted for by pedestrians and bicyclists. Combining this information with the survival probability in such accidents as shown in Fig. 13.8 the significant impact of reducing the allowable vehicle speed becomes obvious.

However such a measure would also have a severe impact on the transportation flow. As the severity of injuries inflicted on the pedestrian or bicyclist is not only a function of the collision speed but primarily of the actual impact harshness of the body on the vehicle, a passenger car should be designed in such a way that it can absorb a substantial amount of an impact, thus inflicting less harm on the other party.

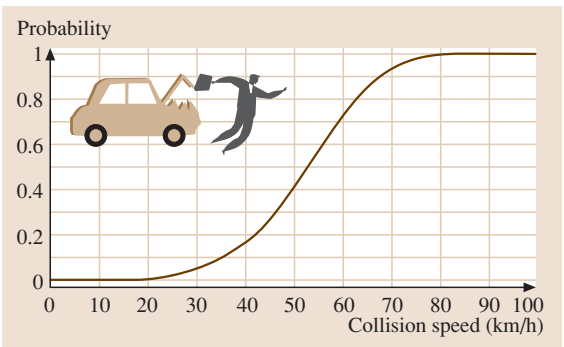


Fig. 13.8 Probability of fatal injury for a pedestrian after collision with a vehicle (after [13.11])

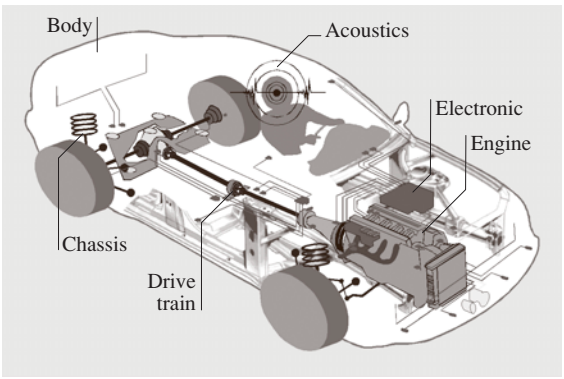


Fig. 13.9 Technological development areas in the vehicle (from ika, Aachen)

13.1.1 Road Transport – Vehicle Technology and Development

The automotive industry is one of the most important global key industries. In order to participate on the

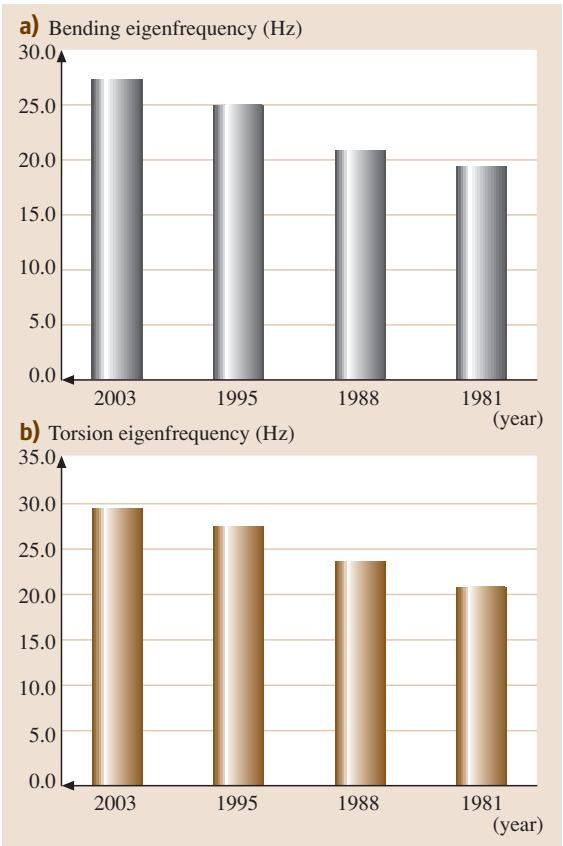


Fig. 13.10a,b Dynamic stiffness of an intermediate-class car model (source: Automobiltechnische Zeitschrift (ATZ)). **(a)** Bending eigenfrequency, **(b)** torsion eigenfrequency

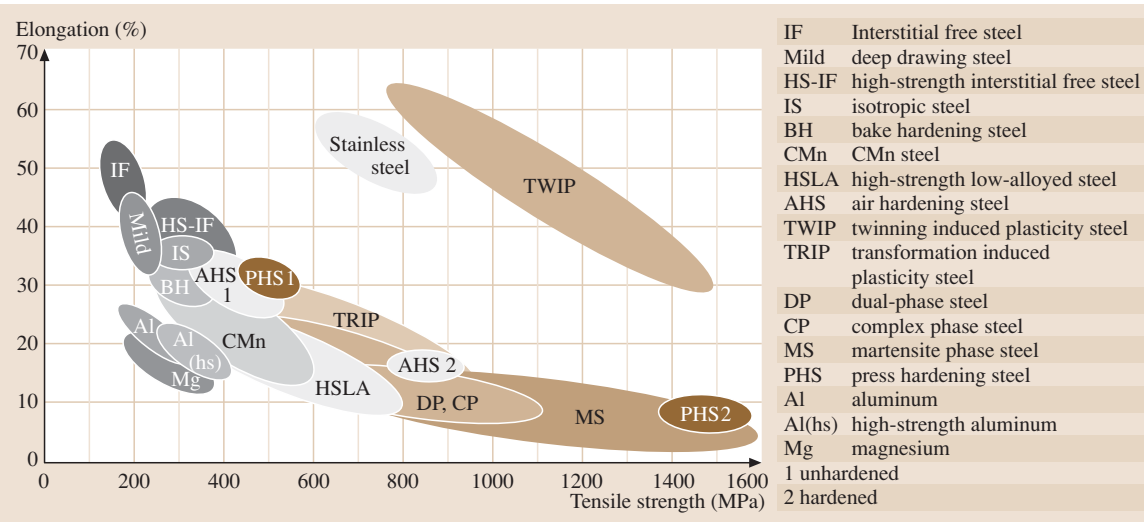


Fig. 13.11 Strength and formability characteristics of various steel grades (source: Forschungsgesellschaft Kraftfahrwesen mbH Aachen (fka))

steady growing automotive market, original equipment manufacturers (OEMs) and suppliers need to generate innovations. Below, the technological developments and the future challenges in the areas of body, chassis, advanced driver assistance systems, and internal combustion engines (Fig. 13.9) are described.

Development Trends in the Field of Automobile Bodies

The body is not only the central load-carrying part of a motor vehicle, it also determines its capacity for passenger and/or freight transport, dominates the outer appearance of the vehicle, and has an important influence on safety and driving comfort.

The requirements of passive safety (noise–vibration–harshness, NVH), driving comfort, and dynamics determine the design of the body structure. In Fig. 13.10, the increasing dynamic stiffness of cars is shown.

Another aspect of vehicle safety of growing relevance is pedestrian protection (Sect. 13.1). The introduction of corresponding regulations in Europe and in Japan has already brought about major changes to the design of the front of many car models. With those regulations becoming more stringent, additional pedestrian safety measures will be launched in the coming years. Moreover, crash compatibility in vehicle-to-vehicle collisions will be an important safety issue for the future, and will probably have a considerable influence on body design as well.

While safety measures tend to result in a mass increase, there is a strong desire to reduce vehicle weights in order to improve handling characteristics, driving performance, and last but not least fuel consumption and CO_2 emissions. However, during the last three decades, the empty weight of passenger cars has increased steadily in spite of intensive efforts regarding lightweight design of body structures. A major challenge in vehicle engineering will be to reduce vehicle weight in spite of additional functions and requirements and cost pressure.

For high-volume cars, the steel unibody is the standard design concept. Substantial improvements in the

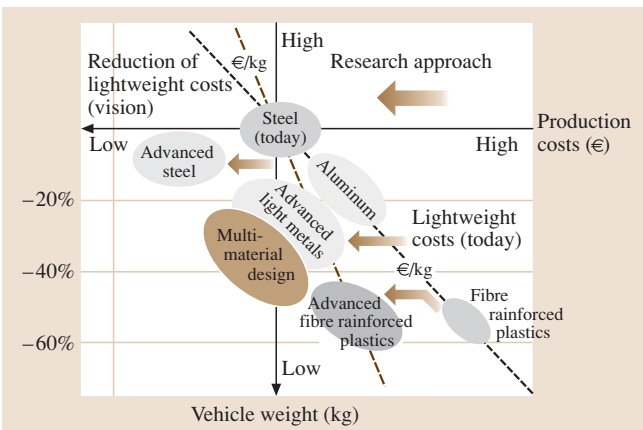


Fig. 13.12 Vehicle weight versus production costs (source: Volkswagen)

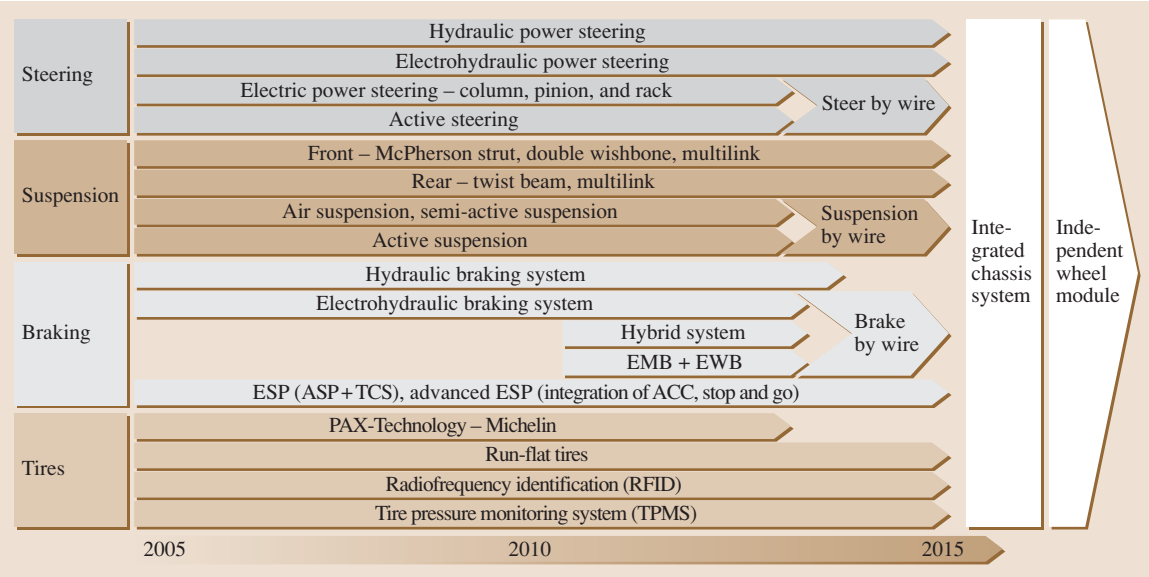


Fig. 13.13 Evolution of chassis components (source: Frost & Sullivan Consulting, 2007)

crash behavior of steel unibodies as well as weight savings have already been realized by the application of new high- and ultrahigh-strength steel grades (Fig. 13.11).

Rolled profiles and hydroformed tubes offering superior strength and stiffness properties in comparison with conventional spot-welded steel-sheet structures. In addition, a trend towards multimaterial design is noticeable. In principle, this means that for each body component or module the most suitable material will be selected, which will result in an intelligent mix of materials such as aluminum, magnesium, plastics, and of course steel. The main purpose is to realize a beneficial combination of low weight and relatively low production costs, as illustrated in Fig. 13.12.

Development Trends in the Field of Chassis

For decades, the trade-off between safety (road holding) and comfort (soft suspension and noise insulation) was the greatest challenge for every chassis engineer. Adaptive components and active chassis systems aim at solving this issue. Over the next 10 years, the portion of electronic systems in the chassis will increase five times in terms of cost compared with today (Fig. 13.13). The more intelligent and controlled systems that are introduced, the more the need for an integrated control strategy increases. The next step is the link to passive systems.

The electronic stability program (ESP), for instance, can be connected with active steering or electric power steering for a combined steering and braking input, or with adaptive air suspension or variable damping systems for roll angle adjustment. Also, ESP sensors can detect the risk of an accident and pretension passive systems.

Regarding suspension types, the McPherson strut and the double-wishbone suspension are the commonly used front suspension types, while the twist beam and multilink types are the choice for rear suspensions. The McPherson strut already has a marked share of 85%; the multilink rear suspension is expected to grow over 55% in the long term.

The developments in brake and steering systems and also damping systems are moving towards the so-called *dry* chassis, which means that hydraulic actuation is replaced by electric motors. This offers more opportunities for control and easier system integration.

Development Trends in the Field of Advanced Driver-Assistance Systems

Due to the increasing traffic volume and growing complexity of driving tasks, technical systems have been developed in order to support the driver and relieve him/her of some of these tasks. These systems are called advanced driver-assistance systems (ADAS). The roadmap of upcoming ADAS is shown in Fig. 13.14. These comfort systems are the basis for safety systems,

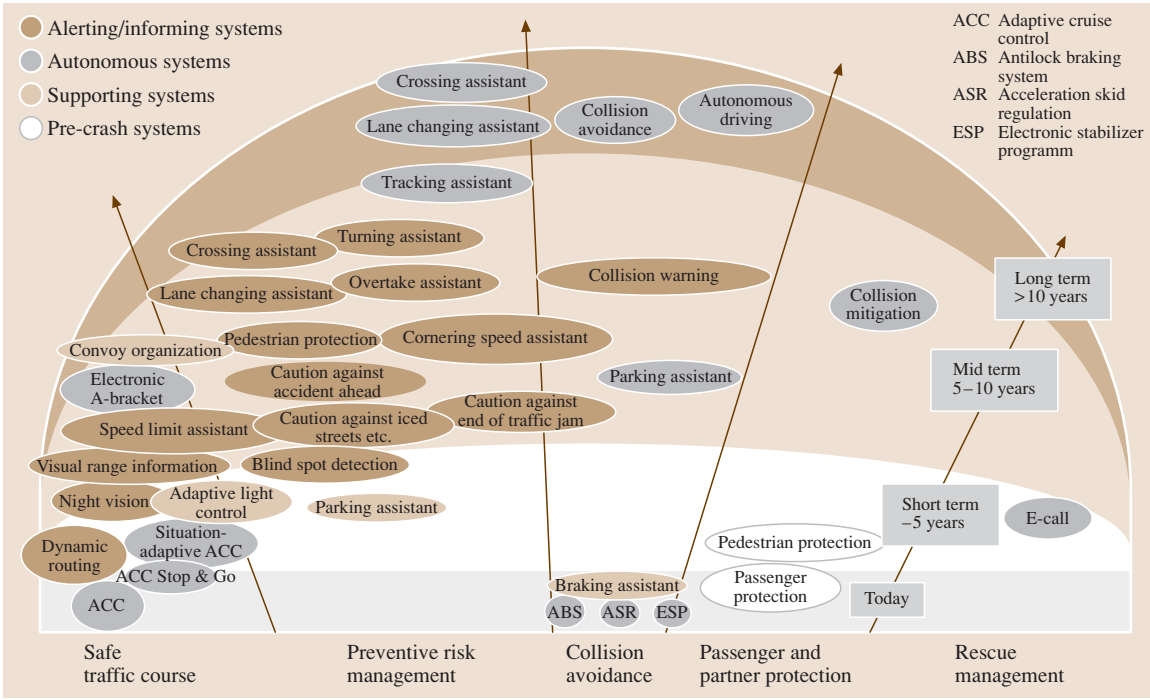


Fig. 13.14 Roadmap ADAS/active safety

because they can avoid accidents or at least reduce the collision speed.

Today the focus of ADAS is increasingly on safety. The European Commission has set the goal of reducing the number of road-traffic fatalities in 2010 by half in comparison with the number in 2001. It is obvious that such a goal can only be achieved with the support of ADAS and active safety systems.

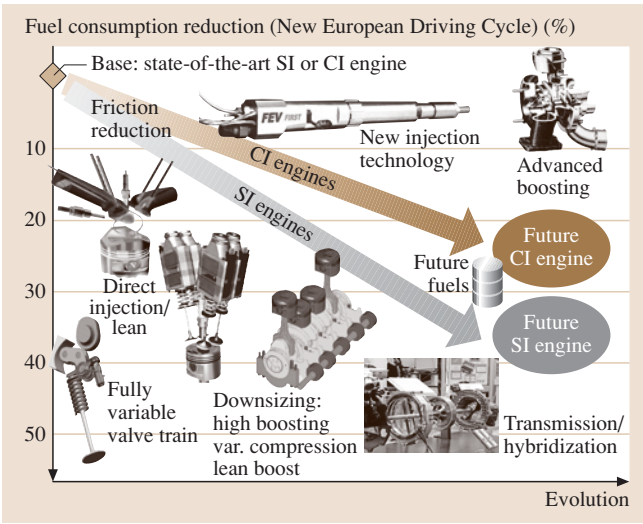


Fig. 13.15 Potential of fuel consumption reduction with power-train technology (source: VKA)

Development Trends in the Field of Internal Combustion Engines

In Fig. 13.15 an overview of current development trends for internal combustion engines (ICE) is given, showing the potential for fuel consumption and corresponding CO_2 reduction of 20–40% for compression-ignited (CI) diesel and especially spark-ignited (SI) gasoline engines.

The main technological trends for future SI engines are direct injection combined with advanced boosting strategies, an increase in variability, e.g., variable valve train, variable compression ratio, and an increasing degree of hybridization. Advanced injection systems (increased injection pressures and higher degrees of variability) and boosting technology are the focus of research and development (R&D) for modern diesel engines. Additional improvements are possible by using future fuels, e.g., tailored fuels produced from biomass, for both combustion concepts.

13.1.2 Aerospace – Technology and Development

Aerospace and Society

As addressed in the general introduction, the wealth of any modern society is based on fast and reliable transport of information, passengers, and freight in an environmentally acceptable way. Aerospace is a mandatory part of the overall transportation system: satellites provide weather information, navigation signals, and transmit messages worldwide; transport aircraft provide fast and safe transportation of passengers and/or freight across borders; and military aerospace is part of any strategy for many years to come. In commercial terms, the future of aeronautics is bright: the increase of passenger traffic has been stable at some 5% per year for three decades, and is forecast to continue this trend for the next decades; for cargo growth is higher, at 6% per year. From Fig. 13.2 it can be derived that passenger transport by aeronautics exceeds that by rail and public transport from 2015 onwards.

In addition, aeronautics has always touched the human imagination: from the story of Daedalus through Leonardo da Vinci to Antoine de Saint-Exupéry. It is an ancient wish of mankind to be able to fly, and space has inspired human thoughts for a long time as well, from ancient cultures, which already knew much about the stars, up to today, where we think about an ever-expanding or finally collapsing universe.

In some fields of activity, aeronautics and space systems merge. After launch space payload carriers not built in space have to cross the atmosphere, coping with gravity and air friction, similar to aeronautic vehicles. More specifically, common interests are in space tourism, hypersonic transport, and rail guns as satellite launchers.

The launch of space vehicles is somehow spectacular, but limited to a few places in the world. Therefore, the public takes quite some interest in such a launch, because it is linked to that dream of mankind mentioned before. On the other hand, and even though aeronautic vehicles are also part of that dream, and are mandatory for modern life, their operation is more heavily criticized. Whereas, for example, railway noise is accepted even in the heart of cities at midnight, aircraft noise seems to be annoying even if hardly noticeable. The public is used to all other kind of traffic, and has accepted their drawbacks, but noise and emissions from aeronautics are more in the focus of today's discussion than their share of the overall transport system. This may be due to the fact that the basis of one's desire to fly is the ease and weightlessness of a bird soaring in the sky. Thus, man reacts more sensitively to aircraft noise and pollution than to pollution of any other means of transport. In reality, much progress has been made over recent decades. Noise has been drastically reduced with the target of keeping the major part of the noise carpet within the airport area (Fig. 13.16), and the amount of exhaust emissions has even been decoupled from the number of aircraft (Fig. 13.17).

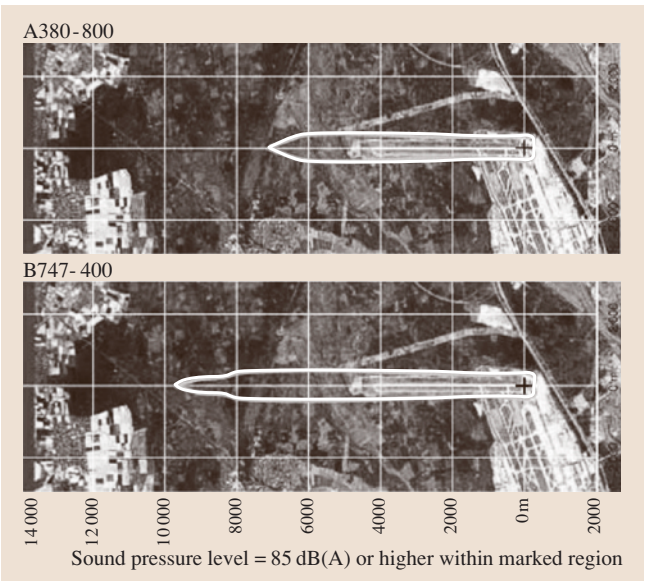


Fig. 13.16 Noise carpet at Frankfurt airport (source: DLH)

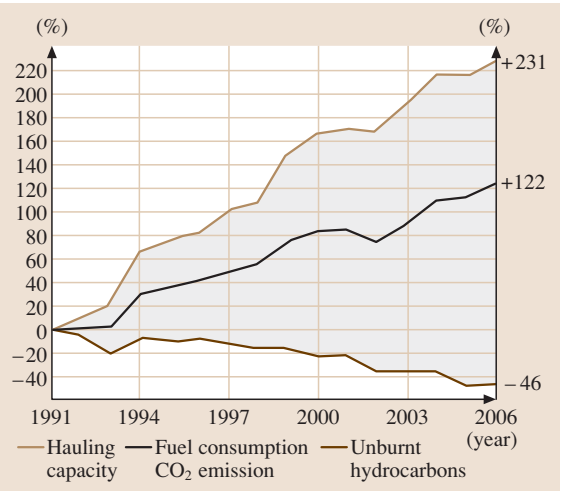


Fig. 13.17 Decoupling of fuel consumption and hauling capacity (source: DLH)

As in other fields, there are many ways to look at aerospace engineering. Firstly, there is the work of specialists in many fields, but one can also consider a system approach combining disciplines into increasingly complex components up to the level of the final vehicles; finally there is a *system-of-systems* approach, when talking about the air transport system (ATS). In astronautics, this system-of-systems approach includes the vehicle, payload, transfer, and ground support. In any case, the total life has to be addressed, from the first flight or launch via planned operation up to the disposal of the vehicle. The importance of the last topic is increasing, in aeronautics due to the issue of global resource management and in astronautics due to issue of space debris.

Aerospace Disciplines and the Design Process

Aerospace in general is an integrated or so-called integration subject, with aeronautics and astronautics as particular fields of activities. Vehicle design is based on inputs from the fundamental engineering disciplines, namely aerodynamics, structures, and systems. As subgroups, flight mechanics, static, kinematics, and others are found but can be allocated to the first three disciplines. In addition, combinations of the original disciplines have emerged, such as aeroelastics and aeroacoustics. Within each of these disciplines, progress has been made in the past, such as new flow control means, new materials, and new electronic systems.

Building upon the foundations laid by work in the individual disciplines, components are created. This is already an interdisciplinary task, as for an aircraft flap aerodynamic performance is needed as well as the structure, including the kinematics and actuators on the systems side. Today, in aeronautics the largest components are the complete aircraft structure and the engine, which are developed and manufactured separately. As for components, their development and final composition into a complete vehicle is a multidisciplinary and interdisciplinary task.

As for most other product developments, aircraft or spacecraft design follows a process which has been set up throughout recent decades. Initially, there is a market analysis, looking for what is missing in the present set of products. In aeronautics, this may be defined by the so-called *transport task*, i. e., the requirement to move a given payload in a given time across a given distance. In astronautics, apart from the vehicle, the payload itself can be the product to be developed, such as a satellite.

From these general requirements set by the market demands, the so-called top-level aircraft requirements (TLAR) can be derived as a basis for the future project office (FPO) work. They will work out a general arrangement (GA) of the vehicle to be designed, working with tools mainly based on statistics and/or simple calculation methods.

This GA will then be given to various departments, such as flight physics or structures. Together with the aircraft programme management, they will start component development, accompanied by tests in wind tunnels, test rigs, or simulators. At this stage, new technologies already available in the respective disciplines can be included in the design process. Increasingly, digital mock-ups (DMU) are used in the first part of the development phase instead of hardware.

With the GA and first performance estimates available, launching customers have to be found. They will influence the final design by their own needs. Depending on the product size and cost, with a certain number of products sold, the actual development starts with the programme *go-ahead*, almost in parallel with the production of the first parts. At the end, specific tests will be carried out on prototypes, which may require some final, hopefully minor, changes in the design. The last

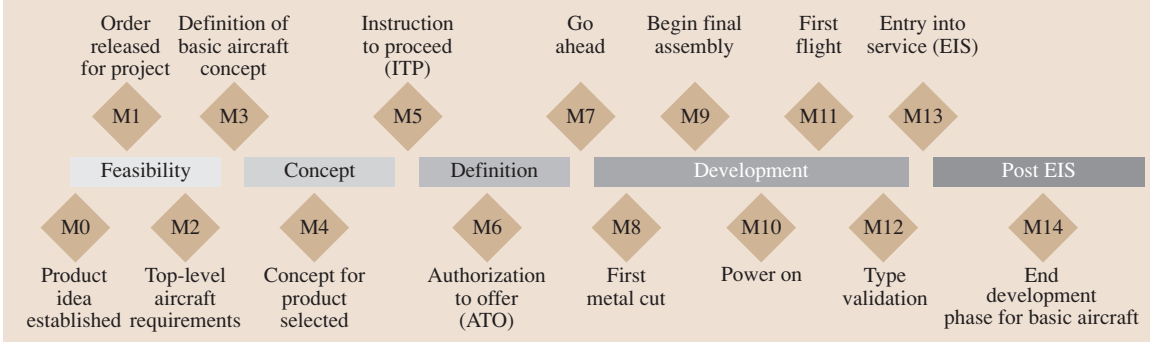


Fig. 13.18 Development process

step is the certification process, which ends with entry into service (EIS).

The whole development process as outlined above is almost the same for any aircraft manufacturer, and can be seen as a series of milestones as shown in Fig. 13.18.

At present, the market analysis may last for 2 years, the predevelopment for some 3 years, the development and manufacture for about 5 years, and the certification for another year. Just before the committed start of a programme, technology development may take place, with technology feasibility or prematurity as a first step, and technology application studies as the final step. Overall, the time between the first thought about a new aircraft and the EIS is approximately 10–15 years.

For example, first thoughts on the Airbus A380 were published in 1989; at that time it was called the megaliner, ultra-high-capacity aircraft (UHCA). Later it became the very large commercial transport (VLCT, as a common Airbus–Boeing feasibility study.

The committed programme start occurred in 2000, and finally the A380 was introduced into airline service in 2007, corresponding to a total of 18years. Of course, all manufacturers try to speed up this process. For the given and already highly optimized standard aircraft configuration, i. e., fuselage, wing, engines, and tailplane as for today's aircraft, further optimization is possible mainly by improving the design chain, including supplier management. However, for a completely new design, such as a blended wing body (BWB) (Fig. 13.19) or an oblique flying wing (OFW), the exhaustive time scale as described above is likely to remain.

Apart from monodisciplinary technologies such as a new material, a new actuator or a specific aerodynamic vortex generator, integrated technologies make their way into the product, too; for example, the fly-by-wire technology was a *must* for the European supersonic transport aircraft Concorde in the 1960s. Without it, it would not have been possible for the pilots to fly this aircraft in all the different flight regimes, in each of which the aircraft control behavior is different. Afterwards in a first attempt, this technology was introduced into the Airbus A310, until it finally became mature in the Airbus A320, which made its maiden flight in 1987. Flow control technology or artificial instability are other examples of technology development, and manufacturing technologies such as friction steer welding or laser beam welding, advanced bonding or surface coating have been developed in the past and are part of the production process today.

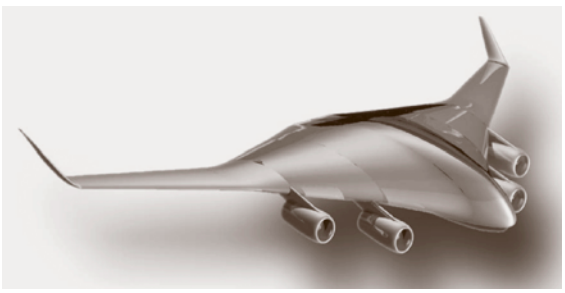


Fig. 13.19 Blended wing body (source: DLR)

Following the complete vehicle assembly, the vehicle will be operated in a larger system. Aircraft navigate with the help of air-traffic control, they are linked to other traffic in the air, and on the ground, especially in the vicinity of airports, they need to be loaded and unloaded and are part of a so-called intermodality concept that links personal and public transport on the ground with air or sea transport.

Challenges in Aeronautical Engineering

Looking at aeronautics' history, it can be structured into three blocks. From its beginning until the end of World War II, physical understanding was the dominating driver. This began with daring pilots in *fantastic flying machines*, permanently hunting for records in range, speed, and altitude. As the next phase, coinciding with the introduction of jet engines, commercial aeronautics emerged. Many different configurations have been studied, such as vertical take-off and landing (VTOL) aircraft such as the Dornier Do 31, supersonic transport in the shape of Concorde, and flying wings such as the Northrop YB-49. This led to the third phase, which is based on today's configuration of a commercial transport aircraft, all looking very much the same regardless of the manufacturing company. This configuration has reached a high level of maturity, so after all the expensive configuration studies, finally civil transport aircraft design and manufacture pays off.

However, with the success of commercial transport, three other issues have emerged. Firstly, airports are increasingly operating at their capacity limits, so it is questionable whether there is any chance to increase air traffic, even if there is a demand for it. Secondly, linked to this, environmental aspects play a leading role. Even though the contribution of aeronautics to global emissions may be small, i. e., at about 2% today, it is debated very intensely. Thirdly, and still linked to growth, safety issues are of increasing importance. Today's reliability rate of 10^{-9} failures per flight hour for critical components will not be sufficient if the number of aircraft

doubles within a decade. In order to reduce the number of accidents in parallel with an increasing number of vehicles, functional hazard analysis must yield greater reliability. This holds true for single components such as an actuator, up to subsystems such as an aileron, and also the complete aircraft system and structure.

Almost at the same time, the European Commission published its *Vision 2020* on aeronautics and the National Aeronautics and Space Adminstration (NASA) published its *Aeronautics Blueprint* on these issues. Both came to similar conclusions: with aeronautics being a vital element for the wealth of society on one side, and the environmental issues linked to it on the other side, in the future there will be additional TLAR, in order to balance transport needs, society needs in terms of safety and security, and environment protection. These additional TLAR may ask for totally different vehicle configurations as well as for new ways of operating these vehicles. In addition, all of these issues ask for a system approach; for example, in contrast to road and rail transport, security will play an increasingly important role in aeronautics. There are an increasing number of studies on the seamless air transport which ask for new ground procedures and probably even new vehicle designs. To define and finally solve all the new TLAR will be a demanding task for any discipline as well as for the overall vehicle and system composition, in aeronautics as well as in astronautics. Therefore, a fourth phase can be expected, aiming for *sustainable growth*.

13.1.3 Rail Transport – Rail Technology and Development

The main difference between railways and other means of transportation is the automatic guidance of the vehicles, leaving only one degree of freedom for a vehicle. This automatic guidance keeps the vehicle on course, defined by the infrastructure. A train can be handled just by controlling the velocity.

However there is no possibility to sidestep other vehicles spontaneously, as the vehicles are guided along the tracks on an accurately defined path. Rail vehicles can only pass each other in specially equipped places. To prevent the system from deadlocks a schedule is absolutely necessary. Together with the timetable there is a need for train protection which prevents two trains heading in opposite directions on one track.

Efficiency, Resistance, and Traction Forces

An advantage of the aforementioned railway schedule is the fact that trains can run almost nonstop from one station to another, as long as the schedule works properly. Thus long and heavy trains can be operated in an energy-saving manner, as they have to be accelerated only once. This small number of accelerations and decelerations becomes even more important regarding a train's ability of acceleration. Because traction forces have to be generated by friction between the wheels and the rail, acceleration forces are limited. The frictional coefficient between rolling steel bodies is always correlated with the relative velocity (slippage) between the bodies (Fig. 13.20).

The maximum coefficient of friction on dry rails is about $f_x \approx 0.45$; if rails are wet or polluted it can decrease to values of $f_x \approx 0.1$. This means for deceleration, where all axles are braked, a maximum deceleration of $b \approx 1\,\mathrm{m/s^2}$ can be guaranteed (13.1). Besides traction power, the maximum acceleration depends on the number of powered axles and the actual friction conditions. For passenger trains this problem is increasingly being solved by using multiple train units in which the traction is distributed over the whole length of the train. Freight trains are still equipped with conventional locomotives, where only four or six powered axles have to pull the entire train

$$F_\mathrm{B} = mb = f_x mg$$
$$\Rightarrow b = f_x g = 0.1 \times 9.81\,\mathrm{m/s^2} = 0.981\,\mathrm{m/s^2}\,. \tag{13.1}$$

Once a train is in motion the low frictional force of the wheel–rail contact offers a big advantage due to the low rolling resistance of steel on steel. Compared with road vehicles, for which rubber is rolling on tarmac, resistance is marginal. Especially for heavy loads this generates an advantage for railways. So when the

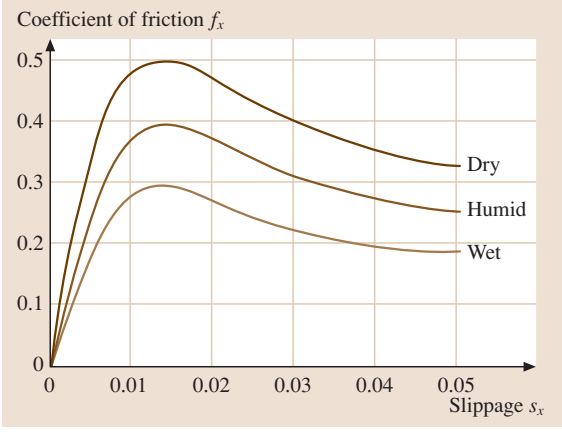


Fig. 13.20 Coefficient of friction

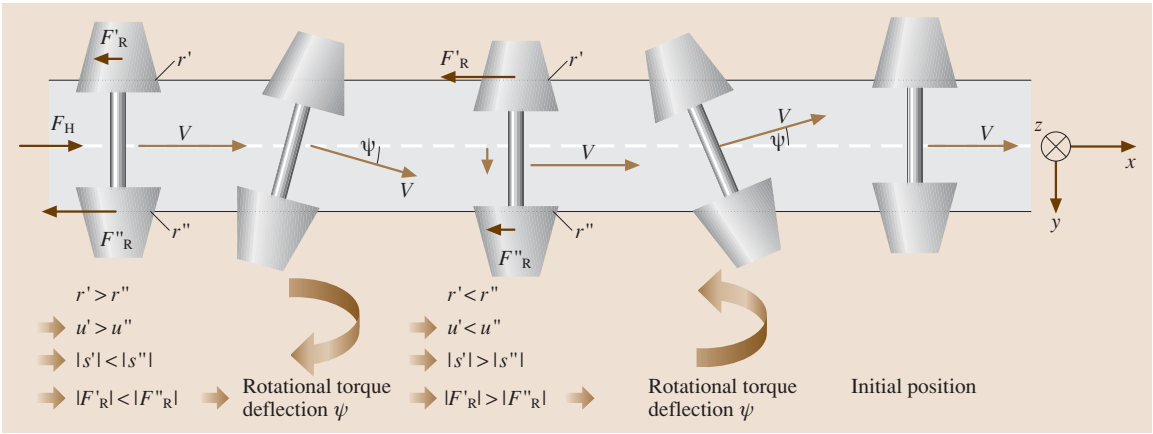


Fig. 13.21 Sinusoidal movement of a (powered) wheel-set (circumferential velocity u > translational velocity v)

train reaches its cruising speed only a small amount of traction is needed to keep the train moving (Table 13.1).

Guidance

Automatic guidance is also an effect of wheel–rail interaction. Conventional wheel sets are guided by an effect called *sinusoidal movement*, which is caused by the wheel profile. In a simplified approach wheel profiles can be regarded as conic, so that deflection of the wheel set (pair of wheels torsion resistant fixed on a shaft) causes different circumferential speeds of both wheels. Together with the slippage–adhesion correlation this leads to different traction forces on both wheels and for this reason to a turning torque acting upon the wheel set.

As shown in Fig. 13.21 the resulting torque on the wheel set always acts in a direction turning a deflected wheel set back towards the centreline of the track. After passing the centreline the orientation of the torque will change and turn the wheel set back into a position offset and aligned to the tracks centreline. At this time the action will start again with opposite sign. Due to the wave-like motion of the wheel set along the track's centreline this is called the sinusoidal movement of the wheel set.

Table 13.1 Rolling resistances of different vehicles $F_R = f_R mg$

Wheel type	f_R	Example of contact pairs
Heavy rail	0.002	e.g., intercity train
Light rail	0.008	e.g., tramway
Tire	0.01	Tire on tarmac
Tire	0.2	Tire on unfortified road
Tire	0.007	Race cycle on tarmac

As well as the desired effect of guidance, the sinusoidal movement causes lateral forces which disturb the riding comfort of the vehicle on the one hand and on the other hand can lead to derailment of the wheel set when a certain velocity is exceeded. To achieve higher speeds and to increase riding comfort damping of the sinusoidal movement is necessary. As sinusoidal movement always coincides with slippage and tangential forces, energy dissipation in terms of noise and wear will occur.

The vast majority of rail vehicles are equipped with wheel sets providing the guidance. However, a few vehicles have single wheels that are independently joined to the axle (Fig. 13.22). Here guidance is achieved by geometrically induced forces. The wheel profile is shaped in such a way that the inclination of the profile increases towards the wheel flange. When a wheel's flange comes close to the rail this will generate a higher lateral share of the normal force between the wheel and rail, which will have a greater magnitude than the lateral force of

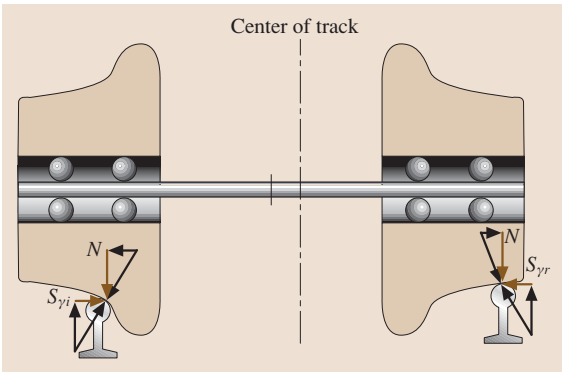


Fig. 13.22 Guidance by single wheels

the opposing wheel, where the flange is far from the rail and thus the inclination of the profile is smaller. Therefore the pair of wheels is pushed back towards the center of the track.

As the single wheels need no axles between the wheels, they are primarily used in low-floor trams to allow an entirely low-floor design without any steps between the wheels. For heavy rail applications only the Spanish manufacturer Talgo provides guidance by single wheels profiting from the fact that single wheels result in smoother guidance on straight tracks due to the lack of sinusoidal movement.

Braking Systems

Because train vehicles are automatically guided the system's safest state is at standstill. Therefore railways are designed to lead to standstill in the case of severe failure. To achieve this failsafe behavior, braking systems play the major part and must themselves exhibit failsafe behavior. In the past after almost every serious accident braking systems were improved to prevent the sort of failure that led to the accident. This in turn led to the current state, where regulations for railway brakes are tight, and render almost any other technical solution other than pneumatic brake systems unworthy.

A main characteristic of the pneumatic railway brake is the main air pipe (MAP), i. e., a pressure line running through the complete train (Fig. 13.23). The pressure within the main air pipe controls the driver's brake valve in such a way that, for detached brakes, the pressure has a maximum value of 5 bar. To apply the brakes the pipe's pressure is decreased and the distributor valve transmits the pressure of the reservoir through to the brake cylinder. Because of this engineering methodology the train brake is failsafe: a loss of energy within the MAP causes the brakes to apply.

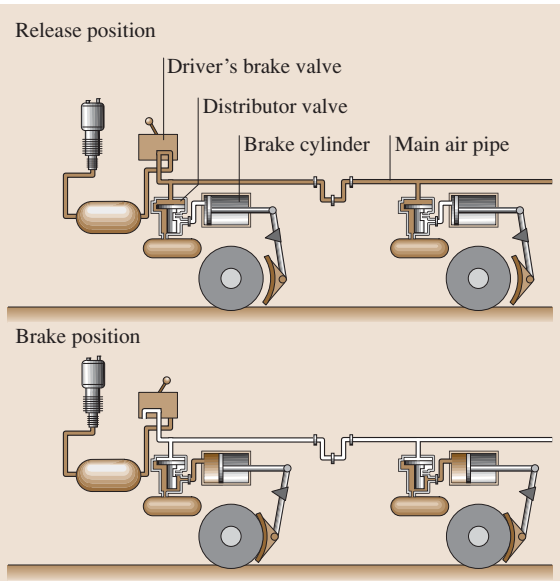


Fig. 13.23 Pneumatic braking system

To convert the cylinder's pressure into a braking torque there are basically two different devices: for ordinary vehicles there is the tread brake, where a brake pad is directly pressed onto the surface of the rail. This on the one hand causes wear and ripples on the running surface of the wheel, which lead to higher noise emissions of the running car. On the other hand damage to the wheel can be caused by thermal energy that has to be absorbed by the brake. For passenger cars braking torque is usually generated by a disc brake, where two brake pads are applied to a disc mounted on the wheel set. The chance of a mechanical damage to the wheel as a direct cause of applying the brake is eliminated. However, damage can still be caused by brakes locking up and forcing the wheel to slide on the rail, resulting in the wheel being planed. So wheel-slide protection (WSP) is needed to avoid wheel flats.

Apart from its failsafe behavior the pneumatic brake has various disadvantages, especially as the response time of pneumatic systems is quite slow. Not only is their controllability for purposes such as WSP poor, but also braking performance suffers from the time it takes the pneumatic signal to reach the last car (in the case of freight trains a delay of approximately 30 s can be expected until the last car has reached full braking performance). To improve the performance of the classical pneumatic brake, passenger trains are usually equipped with additional braking devices such as electropneumatic brakes, in which brake signals are transmitted electronically to achieve better response times. Furthermore powered cars are equipped with regenerative brakes and retarder brakes to increase the efficiency and decrease the wear and thermal stress of the brakes.

Train Protection/Management

As mentioned above a schedule is necessary to protect the system from deadlocks and to ensure that there is no more than one train per track section. To ensure safe operation also in the case of trains not running on time, there is always a system of train protection. Interlocks ensure that no train enters a section that is already occupied by a train. Entry to each section is guarded by a signal that allows or denies access to the block. As a crash is hard to avoid once two trains are on the same

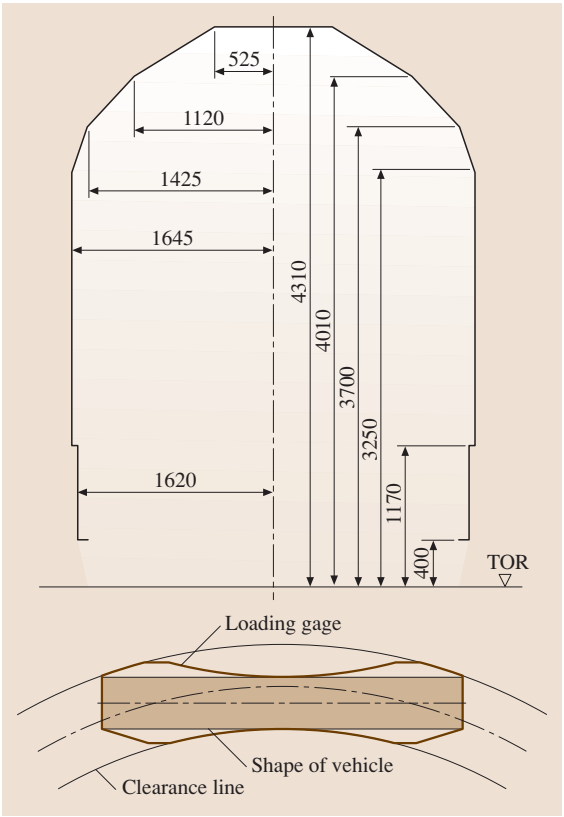


Fig. 13.24 Clearance line for international traffic; shape of vehicle within a clearance line (in mm)

track, railway signals have an even more serious meaning than traffic lights in road traffic. To avoid human mistakes signals are equipped with automatic train protection devices that can stop a train automatically if the driver does not react properly to a signal. These systems are likely to differ from country to country and thus restrain international rail traffic.

Because of their long stopping distances, trains cannot stop within their range of visibility. Thus there are approach signals several hundred meters ahead showing the engine driver what event to expect. As the train has to come to a standstill within the distance of approach and main signal, this distance has to be the stopping distance of the train and is one parameter that defines the maximum speed. Besides the distance of the approach signal the minimum curve radius determines the maximum speed of a railway line. Due to passenger comfort and to prevent cargo from slipping the maximum allowed lateral acceleration on board a train is limited to $1.0\,\mathrm{m/s^2}$.

Structure of Vehicles

The appearance of rail vehicles is primarily dominated by the car body. The size of the car body is determined to a large extent by the infrastructure: the maximum height and width are clearly defined by the clearance outline, and the space beside the track that is guaranteed to be free from obstacles (platforms, bridges, signal posts, etc.). As vehicles also have to pass curves with specified clearance lines, the vehicle's length is also defined by these parameters (Fig. 13.24). A rail vehicle has to be designed in such a way that the vehicle does not touch the clearance line even on narrow curves. So for European standard-gage tracks a typical size of a passenger car has been established with a length of 26.4 m, a width of 3 m, and a height of 4 m. Those cars are usually carried by two bogies with two axles each.

Furthermore the vehicle's mass is limited by the permitted axle load. In Europe, this load is usually limited to 22.5 t per axle, so that a conventional four-axle car has a maximum weight of 90 t. For passenger cars and freight cars with a small payload there are also different arrangements of axles in order to economize tare weight by minimizing the number of axles. Besides the four-axle car the most common type is the articulated train with Jacobs bogies, where two car bodies share one bogie (Fig. 13.25). The disadvantage of this arrangement is that they may only be separated in the workshop, as the link to the bogie replaces the coupling between the cars.

Coupling

The most obvious use of couplings is to create a detachable connection between cars. Furthermore couplings have the function of transmitting longitudinal forces within the train and to prevent the train from being exposed to forces exceeding its specifications. If tensile forces become too high the coupler will break (e.g., if the pulled train is too heavy); if compressive forces are exceeded, for example during a crash, the coupling can

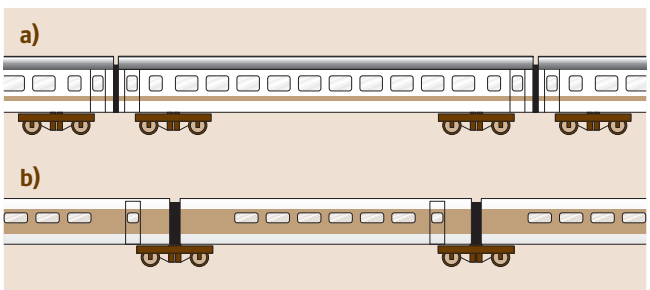


Fig. 13.25 (a) Conventional passenger car. (b) Passenger car with Jacobs bogie

absorb some of the energy. Another important function of a coupling is data transmission. Conventional couplings just transmit the pneumatic signal and energy of the brake's main air pipe. For passenger cars there is also an electric coupling that links the cars' energy supply and various types of signal cables.

In Europe the manually operated screw coupling is standard, offering the possibility to combine all cars without limitation. The advantage of standardization compensates for the disadvantage that the coupling cannot be automated. Automatic couplings in Europe are only fitted to multiple-unit trains in passenger service which have to be coupled solely with other units of similar type. In the USA and in the states of the former USSR automatic couplings are also used for conventional passenger and freight trains.

Shunting of freight trains is – because of the screw-coupling among others – very complicated and requires special shunting yards of huge dimensions. Due to shunting the average speed of a freight car in Germany is below 8 km/h despite the maximum speed of freight trains being 100 km/h loaded and 120 km/h empty.

Within this introduction the significance of today's transportation has been outlined, whereas increasing demand for commercial and recreational freight and passenger travel will enhance the importance of transportation as a highly interdisciplinary field. Transportation and its dependencies offers researchers multiple opportunities to optimize various aspects, i. e., cost, time, safety, and reliability. In this chapter however, the engineering involved in automotive, railway and aerospace will be described in more detail.

13.2 Automotive Engineering

The development of cars differs from the development of other technical products. In this chapter it will be explained why this is the case from a general point of view. Furthermore it will be stated how the topic of *automotive engineering* is imparted.

In principle, a car is a technical product such as a toaster, a refrigerator or a computer. All of these products are things used in everyday life.

13.2.1 Overview

Today every technical device is subject to innovation, development, and production, which are the basics of engineering in general. However, there are differences between cars and their development compared with other technical products, which will now be described.

First of all, a car is a *highly complex* product which has developed tremendously since the times it simply had to transport its load from one point to another. Today, the passenger and their satisfaction are the focus. The passenger is not only kept safe, dry, and warm but also entertained, informed, and even comforted in their seat. The function of transport seems to have become secondary. Still its realization has gained complexity too (Sect. 13.2.2).

Buying a car is normally a *highly individual* process. Thousands of models of hundreds of brands are available. After having decided for a certain brand and a specific model again there are thousands of possibilities to configure one's individual car (at least for premium European car brands).

Furthermore, a car is a *mass product*. Every year, millions of cars are produced and sold in countries all over the world. Characteristics of mass products are: (a) a high number of produced units per year, and (b) there is no direct customer for the development (i.e., there is only a customer for an individual car) but the company itself.

Moreover, cars are *durable products*. About 20 years go by from the first idea to the recycling of cars, and some of the cars remaining at this point even start a second career as classics after this long period of time. In order to achieve this level of durability the development team has to have *a look for the future*: which features will be demanded 3–5 years later? How much is the customer willing/able to pay for the specific feature? Are the development departments of other OEMs working on similar features and what will the specifics of those be? Will there be regulations that might prohibit the use of this feature?

Due to its complexity, required quality, and resulting high development and production costs the car is quite an *expensive* product.

Because of the sum of these characteristics – that the car is an expensive, highly complex, and highly individualized mass product – a lot of *jobs* depend on the automotive industry, not only in many industrial countries. In Germany, for example, one-seventh of all jobs are directly or indirectly dependent on the car industry [13.12]. As a consequence, the automotive industry has a huge influence on society.

To achieve the countless functions and features, design the car and its variants, and set trends in styling a huge *variety of expertise* is needed. In order to bundle this expertise a *well-organized development process* is needed, which again consists of *adapted methods*. Subsections 13.2.3 and 13.2.4 provide details of the characteristics of car development and the methods applied.

As cars are such complex products, a lot of components are *developed externally*, meaning that specialized companies supply their expertise to the OEM to develop components – usually complex, special components such as brake systems (ESP) – and integrate them in association with the OEM into the car context.

Because a car is a product in everyday use, there are thousands of *regulations* to keep the use of it safe and convenient as well as protect the *environment*. The consideration of these regulations again increases the challenge of the generation of functions and features that meet all existing requirements (function, low cost, etc.).

The complexity of a car and its development naturally also affect the costs of development and production. The demanded *high investments* for an entire new car can be up to five billion Euros. Only the aerospace or military industry can boast similar numbers, but have entirely different customers.

Automobiles – A View Back and Ahead

This section provides a general idea of how the car has changed since its invention, what has influenced the development of the car, and on the other hand how the development of cars has influenced our way of life and thereby society and the world. A brief look at the directions that car development might take in the future rounds off this journey.

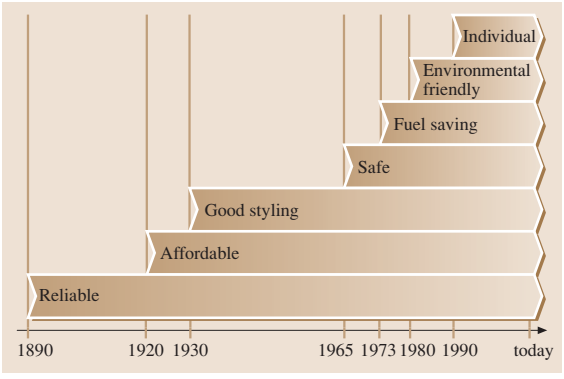


Fig. 13.26 Main aspects of car development and the points in time when they gained in importance

Trends That Have Influenced Car Development. The specifications for a car to be developed include several important points: technical reliability, an appealing styling that makes the customer buy the car, safety for the passengers and other traffic participants, fuel-saving, good economics, and environmental compatibility. Above that, a car also should be individual but nevertheless affordable. An overview of car development over the last 120 years will be given based on these main points.

The first cars built by Daimler and Benz in 1886 were nothing more than coaches with a slow-running Otto engine. These motor carriages had no roof and were not much faster than traditional horse-powered carriages (Fig. 13.27).

One of the first goals of car companies, of which there were increasing numbers in the 1890s, was to achieve greater technical reliability and more comfort for car users. Because no-one had experience in developing cars and no development methods were known, the old-fashioned approach of *trial and error* was used to develop cars (Fig. 13.26).

The Early Car Races as a Field for Testing Inventions. The car companies needed the opportunity to test the inventions they were making and, of course, a forum where they could show off how good their cars were. To do this, car races became popular and helped to accelerate the development of innovations. Car races, for example, accelerated the development of bodies with lower center of gravity, which was good for their driving performance. Hard rally-like races showed how to build a frame that was robust but also lightweight. New techniques such as the electrical ignition system and the

Fig. 13.27 Benz and Daimler carriage (courtesy of DaimlerChrysler archive)

Fig. 13.28 1912 Ford Model-T roadster (copyright 1995–1999 The Henry Ford Organization, Photo: P.833.38916)

shock absorber also demonstrated their reliability in car races.

However, improving the chassis and body was only one way to win races; increasing the engine power was another. In the late 1890s and early 1900s the *classic* approaches for increasing the power of an engine were found: increasing the engine displacement, increasing the compression (which led to the development of higher-quality fuel types), improving the carburetor, including more valves into the cylinder head, locating camshafts above the cylinder head [overhead camshaft (OHC) engines], and so on. Increasing the motor power led to the development of better braking systems for all four wheels which was no natural thing first and optimized suspensions.

Cars for Everyone: the Ford Model-T as an Example. At this stage, cars were still not affordable for most people. Hence, one of the next main goals in car development was to find a way to build cars cheaper so that everyone could buy a car. A pioneer in this field was Henry Ford. He developed a simple car that was cheap to produce and cheap to run – the Ford Model-T (the so called Tin Lizzy, Fig. 13.28).

To achieve this goal, Ford used ideas which are still current: he lowered the producing costs with the introduction of new production methods (e.g., the assembly line), he simplified every component of the car (reaching two goals: greater reliability and lower production costs), and he restricted the variety of options available to choose from.

Good Styling Wakes Desirability. After World War I (WWI), developing more reliable cars was not enough. Increasingly, the customer wanted to have an individual car to stand out from other motorists. This development was accelerated by cars such as the Tin Lizzy; if everyone had the same car, the demand for more individuality and outlining became increasingly important.

Although there was still only one type of carriage – the frame body – emphasis was placed on a number of different body styles: phaetons (Fig. 13.29), tourers, town cars (Fig. 13.30), several styles of convertibles, convertible sedans, Pullman bodies with an extended number of seats, limousines with short and long wheelbase, roadsters, coupés, sedans, etc. The only limit for this kind of technology was the solvency of the customer.

Of course, only very rich people had the means to choose whatever they want from this range, and order a car at an OEM-independent coach builder such as Gläser in Dresden, Castagna or Farina (later Pininfarina) in Italy, Park Ward in England or Saoutchik in France (Fig. 13.31).

In the years following World War II (WWII) cars were optimized in three development dimensions: reliability, affordability, and eye-catching styling. WWII

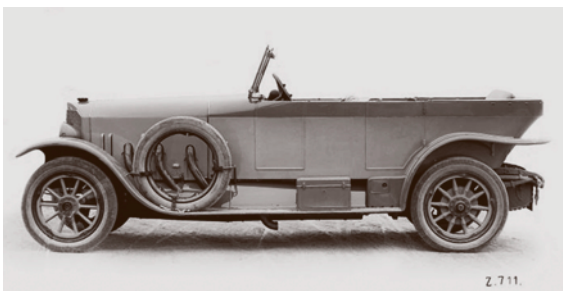


Fig. 13.29 Mercedes 28-95 PS Phaeton (courtesy of DaimlerChrysler)

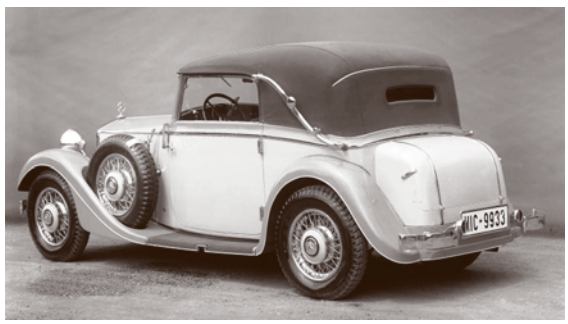


Fig. 13.30 1933 Mercedes-Benz 290 Cabriolet C (courtesy of DaimlerChrysler)

stopped car development throughout the world from about 1941 onwards; car companies changed into defense companies producing tanks, military trucks, jeeps, and even airplanes.

A New Start. After WWII many of the old European OEMs were not able to rise again. Some of them had lost their coach supplier (such as Adler) and were not able to produce. Others had lost their complete production plants as war reparations to the Soviet Union (Auto Union and BMW). Some OEMs had to reestablish themselves (such as DKW and BMW) in Western Germany. All of the cars produced at this time were mainly the same as those produced before WWII.

Car development was thrown back years; the first goal mentioned in this paragraph, technical reliability, became once again the most important. In those days, it was not important to have a good-looking car, but one which was able to fulfill its function to transport people and goods (Fig. 13.32).

As times improved, good design and affordability again became important development goals. A new body style was introduced by Kaiser-Frazer in 1946, the so-called pontoon-type body (Fig. 13.33).

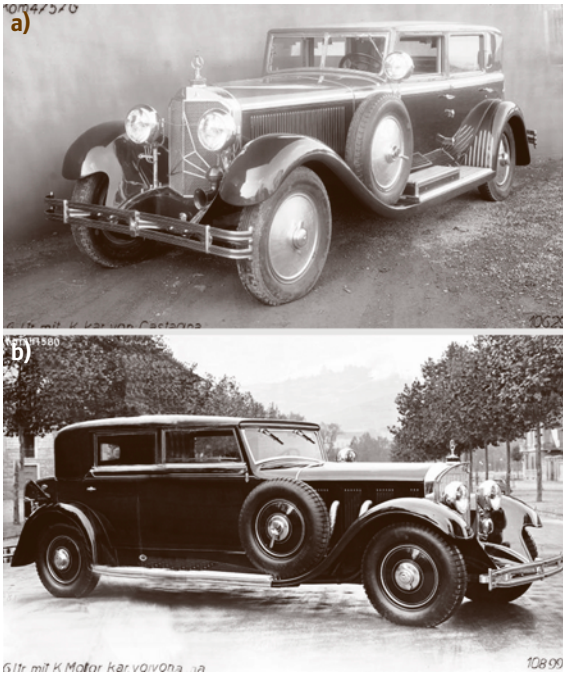


Fig. 13.31a,b Mercedes-Benz Type 630 Modell K, body by Castagna **(a)** and Mercedes-Benz Type 630 Modell K, body by Farina, **(b)** both 1926 (courtesy of DaimlerChrysler)

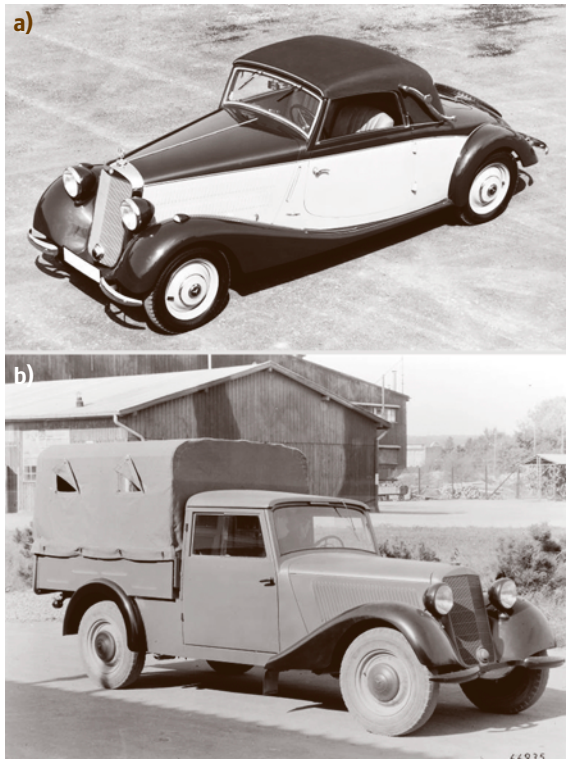


Fig. 13.32a,b Two Mercedes-Benz 170 V; **(a)** 1936 Convertible A, **(b)** 1946 light truck (courtesy of Daimler-Chrysler)

The pontoon-type body was built upon unit body cars, provided more space than the old-fashioned body types, and was much lighter than older designs. On the other hand, this body type strongly restricted the degrees of freedom with respect to individuality of body shaping because of the lack of a separate frame.

Though the idea of the pontoon style and unit body soon showed their superiority, it took almost 10 years

Fig. 13.33 1947 Kaiser-Frazer

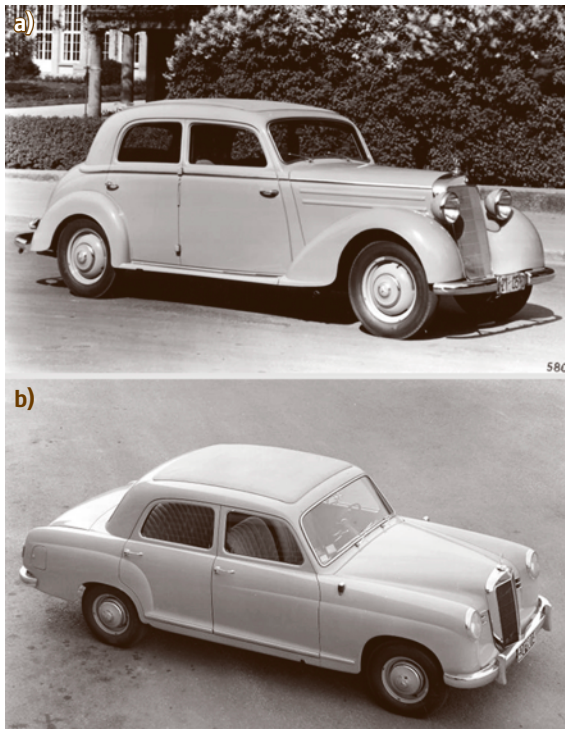


Fig. 13.34 (a) Mercedes-Benz 170 S-V, (b) Mercedes-Benz 180 *Pontoon* (both 1953) (courtesy of DaimlerChrysler)

until the last companies switched their vehicle range. Mercedes-Benz, for example, even as late as 1955 sold a modern and an old-fashioned car in the same class (Fig. 13.34).

Styling Dominates Technical Issues. In the subsequent years the number of car body styles decreased, but styling *exploded*. Some companies, mainly those in the USA and their subsidiaries (such as Ford Germany and Opel) brought out *new* cars every year. The only new thing about these cars was the styling; technical issues were mainly the same – except that almost every car had large tail fins on the rear. The high point of this development was the 1959 Cadillac Convertible (Fig. 13.35).

The Need for Safety Changes Car Development. In the mid 1960s, after the release of a book about car safety (*Unsafe at any speed* by *Ralph Nader*, Pocket Books, New York 1966), the US government took measures to make cars safer: cars had to pass standardized crash tests, which were tightened every 2–3 years. Every car company selling cars in the USA had to pass these tests in order to obtain the approval of the US Department of

Fig. 13.35 1959 Cadillac Convertible (source: http://www.classic-cadillac.com, Classic Cadillac Community, Timmendorfer Strand)

Transportation. Almost immediately all car companies optimized their current car types to pass the government security tests. Research was set up to develop safer cars.

Research was expensive and so was changing current cars to pass new tests. This almost marked the end for many car companies that could not afford this kind of development or had car types in their programs that were not adaptable to the new rules. One class of cars was especially affected by this: the convertibles. The largest market for cars, the US market, stopped buying convertibles because they were considered unsafe (especially in the case of a rollover). As a result almost every OEM that served this market took convertibles out of their program. The casualty of this develop-

Fig. 13.36 1976 MG Midget with US security bumpers (source: http://en.wikipedia.org/wiki/Image:1976.mg.midget.arp.jpg)

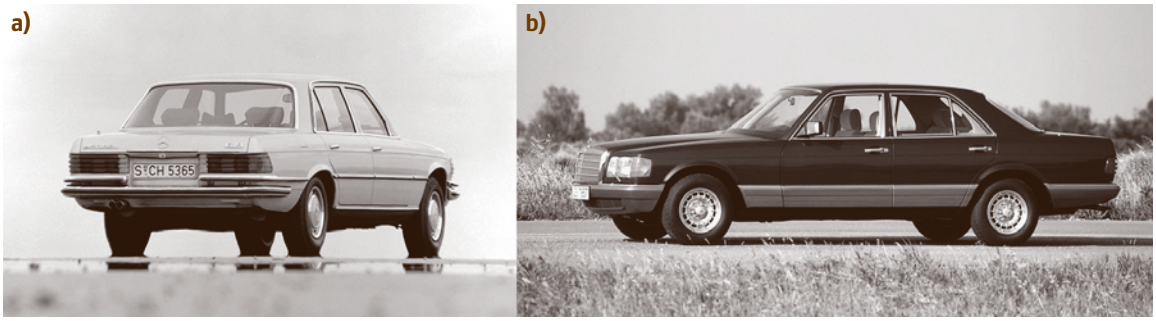


Fig. 13.37 (a) 1978 Mercedes-Benz S-Class, (b) 1979 Mercedes-Benz S-Class (courtesy of DaimlerChrysler)

ment was mainly the British car industry: almost every British OEM had convertibles in their program, including the MGB, the MG Midget (Fig. 13.36), the Jaguar E-Type, the Austin–Healey Sprite, and the Triumph TR-series. All these cars ceased to exist because of these security regulations or were unaesthetically adapted due to the addition of security bumpers and plastic adaptations.

Reliable, Affordable, Good Looking, Safe – and Fuel Saving? In 1973 there was an oil crisis. The oil-exporting nations reduced the amount of oil produced to a minimum and as a consequence the prices of oil and fuel rose tremendously. People could not afford to run their cars in the way they were used to. Now everyone tried to save oil and fuel. In Germany, for example, four car-free Sundays were installed in the autumn of 1973. This naturally also affected the car industry. Several measures were taken by the automobile industry to develop cars that were more economical.

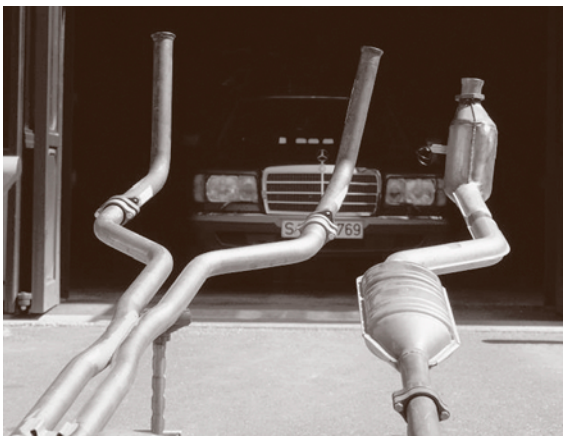


Fig. 13.38 Exhaust system without (*left*) and with (*right*) catalytic converter (courtesy of DaimlerChrysler)

However, now there was a conflict of targets in car development: cars had to be safe, but safety meant that cars became heavier, which resulted in higher fuel consumption. Hence, new ways of developing and producing cars had to be found to solve this conflict. It was not sufficient to concentrate on one development target; for the first time in car history the car had to be looked upon as a unit.

One of the first measures to achieve both security and light weight was to use new materials such as high-strength low-alloy (HSLA) steel and new plastics. The dramatic changes in this area can be seen when compar-

Fig. 13.39 Mazda MX-5 Miata (copyright: Mazda)

ing the 1978 Mercedes-Benz S-Class and its successor (Fig. 13.37).

Plastics replaced chromium-plated metal, engines made out of aluminum replaced those made out of grey cast iron, and aerodynamics was improved. The carburetor was increasingly replaced by the fuel injection, because an injection system controlled by microcomputer was able to reduce fuel consumption in a way that an *analog* carburetor could never achieve. However, there was another reason why the carburetor fell out of favor.

Do Cars Damage the Environment? A phenomenon unknown until that time was experienced from the early 1980s: the dying of forests caused by acid rain, which itself was the consequence of industrial waste gases as well as car exhausts. Measures had to be taken to reduce air pollution.

This trend influenced car development: the catalytic converter for cars with gasoline engines, which had been in use since the mid 1970s in the US, was now used in Europe as well (Fig. 13.38).

The diesel engine was developed for direct injection and common rail technology and, in combination with exhaust-gas turbocharging, a former lame duck named diesel found its way into sports cars, luxury cars, and even convertibles.

Mass Individualization as a New Trend. In the late 1980s the common car changed due to a change in society in the industrial nations: the demand for more individual products. Throughout the 1980s there were several studies for new, small fun cars, but the first one to make it into serial production was the Mazda MX-5 (Miata) in 1989 (Fig. 13.39).

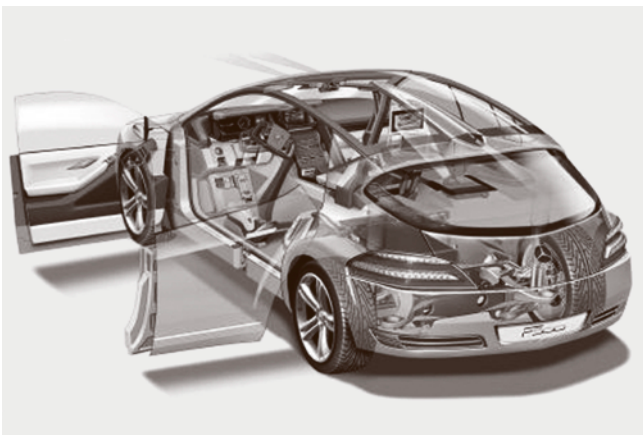


Fig. 13.40 Mercedes-Benz F 500 Mind – a look into automobile future (courtesy of Mercedes Car Group)

Because of the success of this car, other OEMs also created cars for this market: the Porsche Boxster, the Fiat Barchetta, the Mercedes SLK, and BMW Z3 are some examples of this new style.

Another trend which emerged in the 1980s was vans. As they became wealthier, societies in the industrialized nations had more leisure time that was now used for hobbies. For a lot of those hobbies bulky accessories were needed and these hobbies were carried out anywhere but at home – as a consequence cars had to provide more space. The first such vans introduced were the Chrysler Voyager and the Renault Espace in 1984. A classic among these vans is the Volkswagen Microbus.

This mass individualization led to an exploding OEM product range and a trend called crossover, meaning that several car types are crossed. The sports utility vehicle (SUV), for example, is a crossing between a station wagon and an all-terrain vehicle.

What Are the Trends for the Future? Technology is heading towards saving even more fuel. Today new petroleum reservoirs are still being discovered, but it is predictable that one day there will be not enough reserves to run millions of cars let alone other uses.

Steps towards saving fuel include, for example, more widespread use of the diesel engine because it needs less fuel than a gasoline-driven car. Hybrid cars are also an alternative to conventional cars (Sect. 13.2.2). Other technologies under investigation include engines driven by fuel cells, hydrogen, and vegetable oil.

Electronic features for enhancing driving safety, accident avoidance systems, and infotainment systems will probably play an even more important role than today in fulfilling customer demands.

The variety of options and the frequency of replacement of model line versions might be reduced since there is no real return on invest any more.

13.2.2 Automotive Technology

Automotive technology consists of separate areas, which in general have their organizational counterparts in the departmental structure of the automotive company in terms of functional departments and model lines. In this section, an overview of car types and the general targets for car development (and thereby automotive technologies) is given. Then, the technological

areas are briefly described with the demands which have to be taken into account by each and the major functions to be realized.

Car Types

Car types can be classified into one-box, two-box, and three-box concepts.

Cars designed as one-box concepts are perceived as one single volume (one box) and are to be found mainly in the area of vans but also in small cars such as the MCC Smart and Mercedes A-class. Their major advantage is very good space economy, while these cars are comparatively high and thus have disadvantages in terms of aerodynamics and vehicle dynamics due to the high center of mass.

Two-box concepts divide the car into two different volumes, the front volume usually used for the engine bay, and the second volume for passenger compartment and/or loading space. Two-box concepts are used for SUVs, station wagons, and cars in the compact class such as the Volkswagen (VW) Golf.

Finally, the three-box concept is the classic division of the car body into three volumes, separating the engine bay from the passenger compartment and the trunk. This concept is used for limousines, classic coupés, convertibles, and roadsters.

Targets for Car Development

Customer Demands. Customer demands are the most crucial targets for car development. Methods for the translation of customer demands to technical requirements are described in the literature, namely the *house of quality* [13.13].

When clustering the demands that customers and society place on cars one can identify various target conflicts. What is needed is cars that are:

- Safe
- Emotional
- Comfortable
- High quality
- Highly reliable
- Low emission
- Low noise
- Highly recyclable
- Have decent driving performance and load capacity
- Low cost, both in acquisition and utilization

A rough overview of the major design requirements in different market segments is shown in Table 13.2.

Regulations. Regulations from legislation, car associations, and consumer groups are a major influence on car development and the final products themselves. In different countries, different regulations have to be taken into account to:

- Be able to introduce a car model (line) into this specific domestic market at all
- Be able to fulfill the consumer expectations, especially concerning car safety and fuel economy

The European Union has published far more than 50 regulations concerning active and passive car safety (prevention of accidents, reduction of consequences of accidents, car emissions, etc.).

Causes for the definition of regulations are manyfold: from the protection of the domestic markets and domestic car manufacturers to protection of people, the environment, and the traffic.

Table 13.3 shows as an example the relevant standards concerning safety published by the National Highway and Traffic Security Association (NHTSA), an operating unit of the Department of Transport (DoT) of the US government.

Technical Requirements – Specific Example: Climate Stress. Technical requirements for car development are derived from a number of external and internal influences on the car and its operation. Here, as an example area, climate parameters to be taken into account when designing a car, are described.

A car and its components are exposed to a multitude of external factors in terms of climate stress. The strength of these factors depends on the planned operating area of a car and has to be taken into account when designing a car:

- Temperature: the maximum and minimum operating temperature and the temperature levels where the car and its components are kept (e.g., during transport).
- Humidity: from extremely dry conditions (e.g., the Mojave Desert) to tropical conditions – the most stressful environment is a humid, hot climate.
- Water: rain, car wash (with different kinds of washing equipment).
- Sand and dust: a challenge for the layout of sealings (e.g., keeping the passenger compartment free of dust, and taking care that the air convection compo-

Table 13.2 Design requirements in different market segments [13.14]

Mini car	Small car	Family car	Luxury car	Performance sedan	Sports car/coupé	Rallye or track car
Versatile accommodation	Versatile accommodation	Versatile accommodation	Distinctive style	Fast appearance	Sleek styling	Priority to function
Small frontal area	Small frontal area	Low C_d and reasonable area	Low C_d	Low C_d	Low C_d	Ground effect and low C_d
Small engine	Small engine	Choice of engines with diesel option	Wide choice of engines, optional diesel, possibly with turbo	Large engine with fuel injection and/or turbo	Large engine with fuel injection and/or turbo	Maximum power output
Good performance	Good performance	Good performance	Smooth performance	Fun to drive	Fun to drive	Quick response, ultimate handling
Maximum fuel economy	Good fuel economy	Good fuel economy	Good fuel economy for class	Good fuel economy for class	Performance first	Performance first
Low cost	Low cost	Low cost	Value for money	Value for money	Cost secondary	Cost no object
Ride secondary	Adequate ride	Good ride	Good ride	Good handling	Good handling	Maximum road holding
Easy to service and repair	Easy to service and repair	Easy to service and repair				Fast repairs at p.t.o. service stops
Minimum mass	Minimum mass	Minimum mass	Controlled mass	Good power-to-mass ratio	Good power-to-mass ratio	Low mass
Maximum package for size	Maximum package for size	Maximum package for size	Maximum package for size	Reasonable luggage room	Reasonable luggage room	To carry long-range fuel tank and large-section spare tire
Some noise acceptable	Reasonable noise	Low noise	Very quiet interior	Quiet at high speed	Quiet at high speed	Noise not important
4 seats	4 seats	5 seats	5 seats	2+2	2+2	2 seats only
FWD	FWD	FWD or RWD	FWD or RWD	FWD or RWD	RWD	RWD

nents in the cooling system do not become covered with dust.

- Sun: especially for lower parts in the cockpit, high temperatures up to 80 °C and above lead to thermal stress and aging.
- Corrosion: especially for operation on salted winter roads and in coastal climates.
- Chemical fluids: engine oil and fuel can come into contact with certain areas of the car.
- Air pressure: extreme operating conditions in high-altitude areas have to be taken into account when designing seals and membranes.

Body

Targets for body design can be divided into two different classes: targets relevant from the point of view of the end customer and those relevant for internal optimized production [13.15].

Table 13.3 Prescriptions of the federal motor vehicle safety standards (FMVSS) (relevant extract) [13.16]

Standard No.	Title
101	Controls and displays
102	Transmission shift lever sequence, starter interlock, and transmission braking effect
103	Windshield defrosting and defogging systems
104	Windshield wiping and washing systems
105	Hydraulic and electric brake systems
106	Brake hoses
108	Lamps, reflective devices, and associated equipment
109	New pneumatic bias ply and certain specialty tires
110	Tire selection and rims
111	Rear view mirrors
113	Hood latch system
114	Theft protection
116	Motor vehicle brake fluids
117	Retreated pneumatic tires
118	Power-operated window, partition, and roof panel systems
119	New pneumatic tires for vehicles other than passenger cars
120	Tire selection and rims
121	Air brake systems
124	Accelerator control systems
125	Warning devices
129	New nonpneumatic tires for passenger cars
135	Light vehicle brake systems
138	Tire pressure monitoring systems
139	New pneumatic radial tires for light vehicles
201	Occupant protection in interior impact
202	Head restraints
203	Impact protection for the driver from the steering control system
204	Steering control rearward displacement
205	Glazing materials
206	Door locks and door retention components
207	Seating systems
208	Occupant crash protection
209	Seatbelt assemblies
210	Seatbelt assembly anchorages
212	Windshield mounting
213	Child restraint systems
214	Side impact protection
216	Roof crush resistance
219	Windshield zone intrusion
223	Rear impact guards
224	Rear impact protection
225	Child restraint anchorage systems
301	Fuel system integrity
302	Flammability of interior materials
303	Fuel system integrity of compressed natural gas vehicles
304	Compressed natural gas fuel container integrity
305	Electric-powered vehicles
401	Interior trunk release
403	Platform lift systems for motor vehicles
500	Low-speed vehicles

Targets relevant for customers are:

- Appealing design
- Maximum safety
- Minimal fuel consumption
- High comfort
- High level of functionality
- High quality and long life time
- Attractive/acceptable price
- Low maintenance cost
- Low noise emissions
- Usable every day

Targets relevant for production are:

- Easy assembly
- Utilization of existing production machinery
- Small number of different parts
- Easy to manufacture
- High, constant process quality

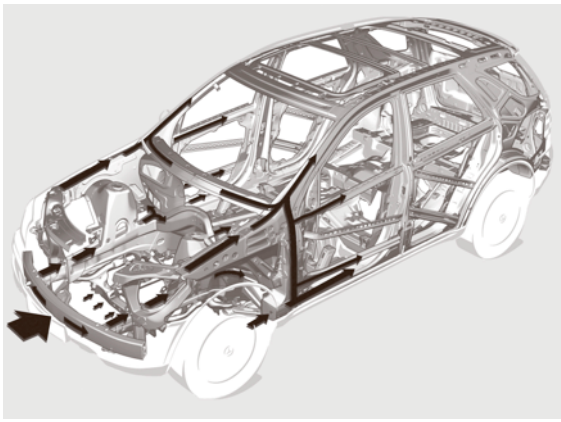


Fig. 13.41 Visualization of the load paths in the car body (courtesy of Mercedes Car Group)

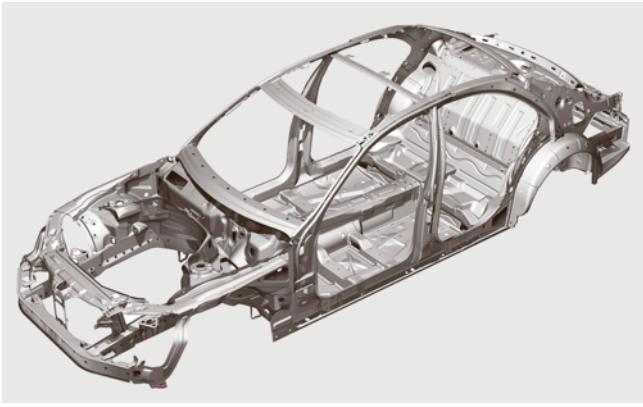


Fig. 13.42 Car body of a limousine (courtesy of Mercedes Car Group)

- Carryover of part and platform strategies
- Optimized utilization of material
- Low production cost

The basic layout of the body is influenced by the basic load cases of a car: bending case, torsion case, combined bending and torsion, lateral loading, and fore and aft loading [13.17]. These load cases lead to stresses in the vehicle body structure. Theses stresses must under the worst load conditions be kept at acceptable limits. The torsional and bending stiffness of the vehicle structure is a main influence on the NVH performance of the car, the tightness of sealings, and high-speed performance, especially of convertibles [13.18–20].

There are various basic concepts for the layout of body types. The ladder frame, historically the first and most flexible way to build car bodies, had its disadvantages especially in terms of weight. Thus, today in most cases the car body is an integral structure with sheet-steel parts spot-welded together (Figs. 13.42, 13.43), providing structural and other functions. Since integral structures – also due to their flexibility in terms of design and utilization of different materials (Fig. 13.40) – are very complex, traditional mechanical analysis cannot be used to predict their behavior in the layout phase, but rather finite element (FE) analysis must be used to determine the optimal shape and material usage. This is also true for crash simulations of the body (Fig. 13.41).

The exterior of the body is defined by three major influences: styling, aerodynamics, and packaging. Together with the body design itself, these are the tightest simultaneously running constraints in the early phase of car development.

Because weight reduction while optimizing stability and crash performance is a challenge which cannot be met by using conventional steel bodies, lightweight design, e.g., using aluminum or tailored blanks, is used to optimize weight, stiffness, and crash performance while trying to limit the production cost.

Since sheet-metal stamping of body shell parts for the car assembly process is partly outsourced, in the car body development process, modules are defined which are delivered by suppliers preassembled and then connected as a whole to other body parts. A typical example is the front bumper, where often, besides the body sheet metal parts, electrical components such as the front lights and the ventilation device as well as crash deformation elements made from plastic material are integrated into one module that is delivered just in time and just in sequence to the production line of the car manufacturer.

Car body properties are the significant parameters used to describe the behavior and performance of the car body. Tolerances in the design process as well as in the assembly process contribute a lot to the body quality and the perception of car quality by the customer. Too large or too inhomogeneous joints between two body sheet metals deliver a sense of low quality in engineering and styling.

Chassis

Overview. Targets for chassis design from the point of view of the end customer are related to the class of vehicles (buses, trucks, SUV, convertible, etc). The most fundamental differences among the requirements for different classes of vehicles are between passenger and

commercial vehicles [13.14]. For passenger vehicles, the major concerns are:

- Ride comfort
- Good handling characteristics (depending on the style of driving)
- Provision of these features over a wide range of different driving scenarios and scenarios

For commercial vehicles, the driving force is economical operation. Thus, design is usually based on a fully loaded vehicle being the most economic way to transport goods from point A to point B, and long time of operation. The major concerns here are:

- Low cost
- Reliability of operation

Between these two different classes of vehicles with completely different demands in terms of function and performance are the requirements for buses, building a compromise between these classes.

The function of the chassis is to isolate the passengers and/or load from shocks and vibration caused by the roughness of the driving surface. Two basic components are used to cover this function:

- Springs, providing flexibility
- Dampers, absorbing energy

Figures 13.44–13.46 show different car axles.

The basic components of the chassis are shown in Fig. 13.47 as a design principle.

The driving characteristics of a car are dependent on a number of chassis and car concepts and layouts. The concepts differ in terms of the position of the engine inside the car (front engine, mid engine, rear engine) and the type of wheel drive (front-wheel drive, rear-wheel drive, and four-wheel drive).

Most of the disadvantages of the car concepts described briefly in Table 13.4 are reduced nowadays through the application of electronic driving support systems.

For further reading on automotive chassis in general, refer to [13.21–30].

Brakes. Brakes are among the most important components in a car. They have to operate under any circumstance, and under every operating condition have to provide the functionalities to:

- Decelerate the car in a controlled and repeatable fashion and, when appropriate, cause the car to stop
- Permit the car to maintain a constant speed when traveling downhill

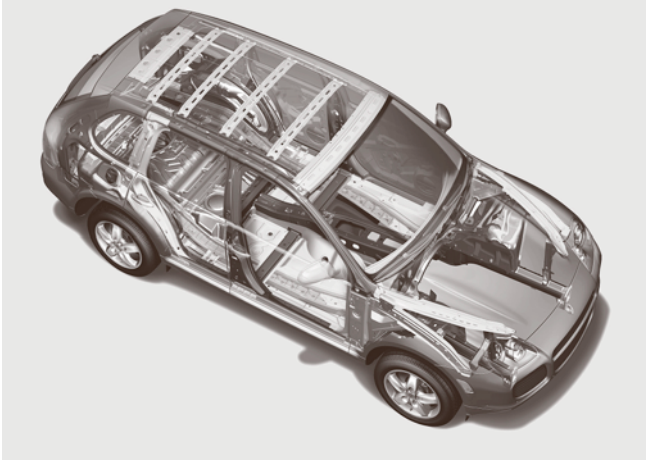


Fig. 13.43 Car body of a sports utility vehicle (courtesy of Dr. Ing. h.c. F. Porsche AG)

- Hold the car stationary when on the flat or a gradient [13.31]

These functionalities have to be provided:

- On slippery, wet, and dry roads
- On rough and smooth roads
- On split friction surfaces
- During straight-line braking or when braking on a curve
- With dry and wet brakes

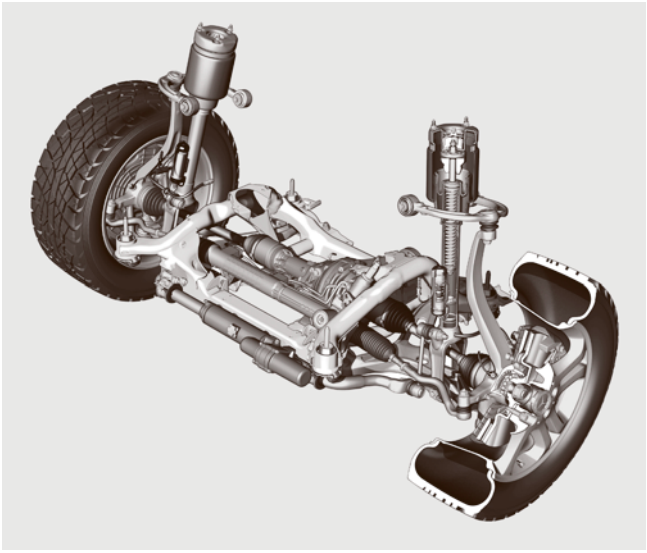


Fig. 13.44 Car axle of a sports utility vehicle (courtesy of Mercedes Car Group)

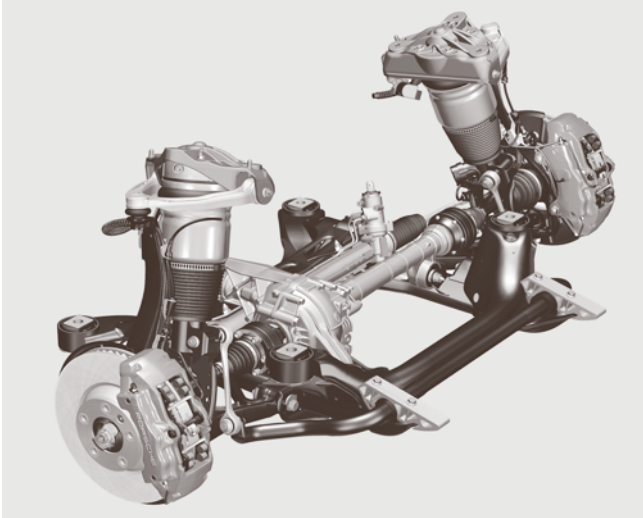


Fig. 13.45 Car axle of a sports utility vehicle (courtesy of Dr. Ing. h.c. F. Porsche AG)

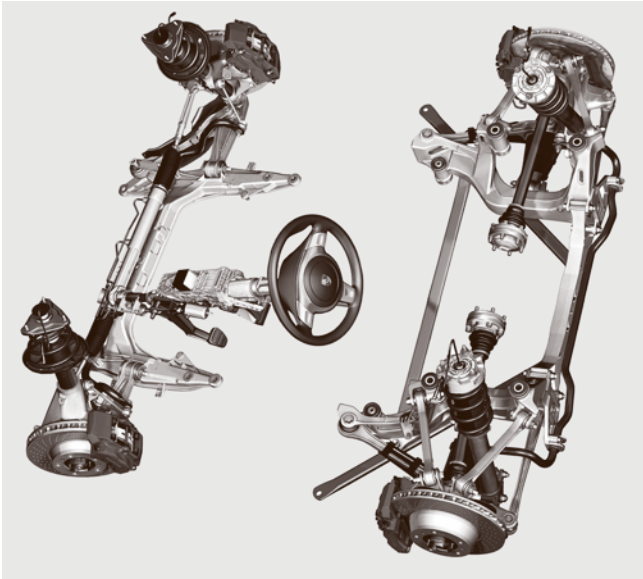


Fig. 13.46 Front and rear axle of a sports car (courtesy of Dr. Ing. h.c. F. Porsche AG)

- With new and worn linings
- With a fully loaded or unloaded car
- When the car is pulling a trailer or caravan
- During frequent or infrequent applications of short or lengthy duration under different temperature conditions and speed levels
- During high and low rates of deceleration

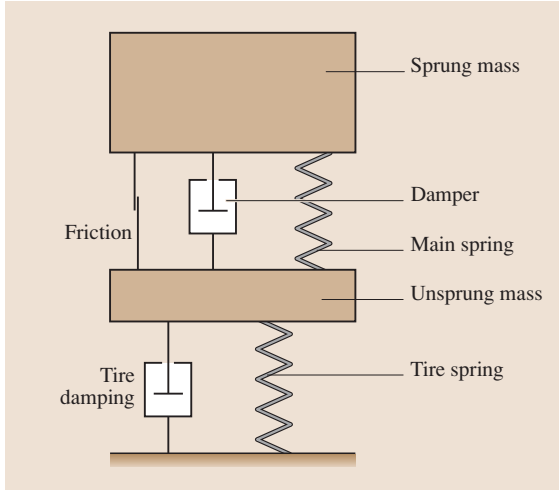


Fig. 13.47 Design principle of a suspension [13.14]

- When used by both experienced or inexperienced drivers

Each car is equipped with two independently operating brake systems, thereby providing redundant functionality at each point of time.

Comfort functions, safety functions, and even functions influencing other car dynamics have been implemented in braking system over time. Figures 13.48–13.50 show the major milestones in the development of car brakes.

Today's hydraulically operating brake systems consist of the brake pedal assembly, the brake booster amplifying the pedal force when the brake pedal is pressed down, the master cylinder initiating and controlling the process of braking by providing the necessary hydraulic fluid pressure, the regulating valves necessary to implement functionality such as antilock braking, and the control software enabling high-level control of car dynamics using brake functionality such as ESP.

Modern brake technology is often derived from insights found in race applications [13.32].

For reasons of comfort, brake noise is a major issue in the design of today's braking systems. Although intensive theoretical and experimental research has been undertaken over many years, the mechanism of brake screech is not yet fully understood, and this problem is one of the most common reasons for warranty claims on new vehicles [13.20, 33–35].

Power Train

The main requirements for the power train in a car are:

Table 13.4 Car concepts in term of placement of engine and type of wheel drive

Placement of engine	Type of wheel drive	Major benefits	Major disadvantages	Sample cars
Front engine	Front-wheel drive	Tame driving characteristics, less design effort	Tight packaging in engine bay, good traction only if light load is carried	VW New Beetle Mercedes A-class
Front engine	Rear-wheel drive	Good packaging restrictions	Weight distribution not optimal	Mercedes C-/E-/S-class
Mid engine	Rear-wheel drive	Good weight distribution, excellent driving dynamics	Not applicable in passenger cars (>2 seats) due to packaging constraints	Porsche Cayman
Rear engine	Rear-wheel drive	Good traction, excellent take-off characteristics	Cooling costly, directional stability difficult to cover by design	VW Beetle Porsche Carrera

- Accelerate the car from standstill up to an arbitrary speed
- Allow for quick regulation of torque and engine speed for good dynamic behavior
- Provide good efficiency in terms of space consumption inside the car as well as energy efficiency and low mass of the components used
- Ability to deal with shocks and extreme temperature conditions

The power train consists basically of the engine, transmission, and clutch.

Engines. In automobiles, 3-, 4-, 5-, 6-, 8-, 10-, and 12-cylinder engines are used as a straight engine, a V-engine, a VR-engine, a Boxer engine or a W-engine [13.15]. Current trends are for diversification of engine concepts also in terms of size and power output, from small engines aimed at drastic reduction of fuel consumption to large engines with high power output for ultimate and comfortable performance.

Combustion Engine. The combustion engine is the most commonly used engine in passenger cars. The working principle is to convert gasoline into motion, usually on the basis of the Otto cycle. The four strokes of the Otto cycle are:

- The intake stroke, during which air enters the combustion chamber
- The compression stroke, during which the air in the combustion chamber is compressed
- The combustion stroke, during which a spark of gasoline is added to the compressed air, ignited, and the air–gas mixture explodes

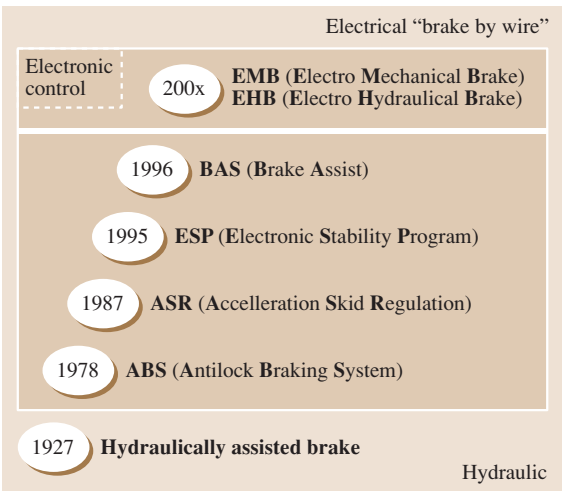


Fig. 13.48 Major milestones in the development of brake systems [13.36]

- The exhaust stroke, where the exhaust fumes leave the combustion chamber

Diesel Engine. Characterized by compression ignition of the air–diesel mixture, the diesel engine has become a favorite type of propulsion system in some European countries due to low fuel cost and tax reductions. The advantages of the diesel engine compared with the Otto engine are:

- Better efficiency, especially in the part-load range, resulting in
- Lower specific fuel consumption

- Lower emissions of carbon dioxide and carbon monoxide
- Diesel fuel is easier to manufacture
- High reliability and long service life

Disadvantages are:

- Higher output of nitrogen-oxides emissions and soot particles
- Higher weight
- Lower engine speed
- Less smooth engine running, and less impulsive reaction to driver signals

Future diesel engines will have to meet the following demands:

- Good environmental properties (e.g., use of particle filters)
- Further reduction of fuel consumption (e.g., increased injection pressure)
- Reduction of production cost (e.g., use less expensive material or simpler engine concepts)
- Longer lifetime
- Even greater reliability
- Greater comfort (reduction of vibration, smooth torque development over engine speed)
- Reduction of noise emission

Besides the combustion engine (p.t.o. or diesel), which is by far the most commonly used propulsion unit in today's cars, there are a lot of different engine types which are used very rarely or in research prototype cars, e.g.:

- The electrical motor, powered by battery, fuel cell or by a generator operated by a combustion engine (Figs. 13.51–13.53)
- The Stirling engine, which has very low emissions and is characterized by a smooth and continuous torque output even at low engine speed
- The gas turbine, which is characterized by low vibrations and low emissions

For further information on automotive engines in general, refer to [13.37–46].

Low-Emission Engine Concepts. The rising number of cars, especially in urban areas such as Los Angeles, CA, has led to the definition of regulations aiming at reduction of air pollution and fuel consumption. Zero-emission vehicles (ZEV), today only effectively realizable using electrical motors or fuel cells, and ultralow-emission vehicles (ULEV), are required to achieve these targets. Today, a certain percentage of all cars sold in the market of California have to meet the restrictions of ZEV/ULEV, otherwise the automotive manufacturer has to pay penalties.

One contribution to ULEV are cars with hybrid drives (Fig. 13.54). These drives consist of a minimum of two different energy conversion units and use two different energy storage methods for means of vehicle propulsion. The main potential benefits of the hybrid drive are:

- Reduction of fuel consumption
- Reduction of emissions
- Reduction of noise

There are three different concepts for hybrid drives: parallel, serial, and mixed.

Serial hybrids use electrical energy for wheel drive, and the energy stored in the battery is recharged from

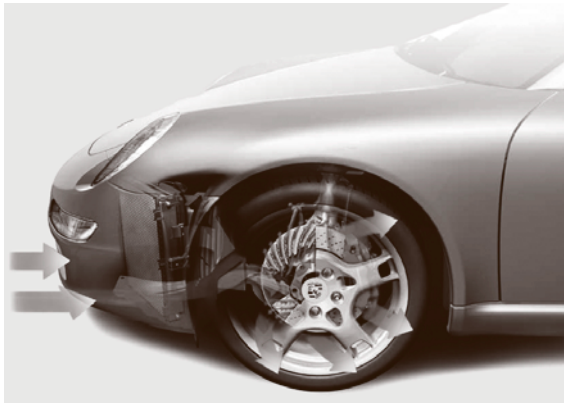


Fig. 13.49 Conduction of airstream for cooling of brakes (courtesy of Dr. Ing. h.c. F. Porsche AG)

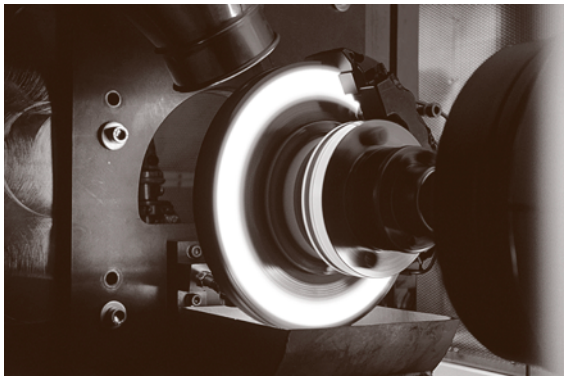


Fig. 13.50 Brake test on a testing machine (courtesy of Mercedes Car Group)

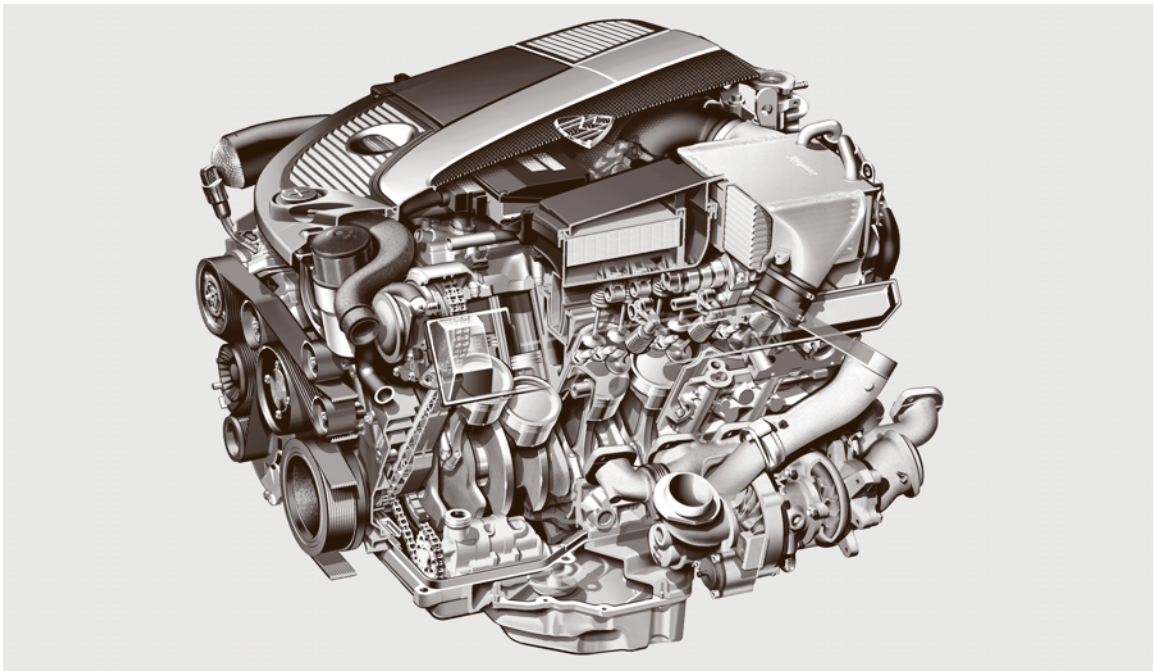


Fig. 13.51 Twelve-cylinder combustion engine (courtesy of Mercedes Car Group)

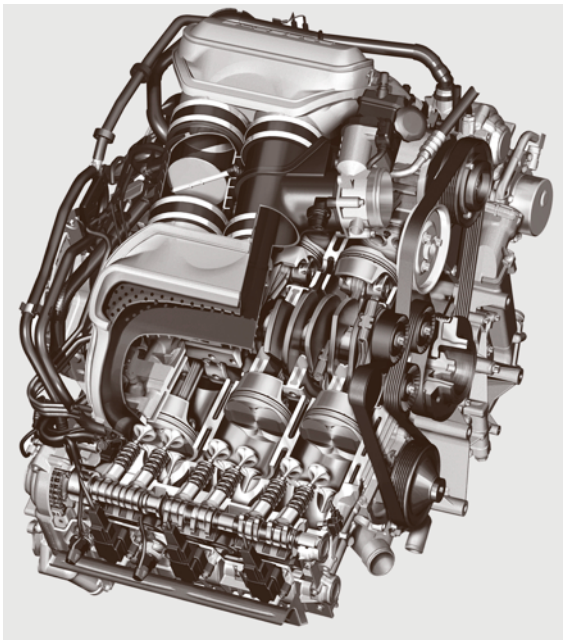


Fig. 13.52 Six-cylinder combustion engine (courtesy of Dr. Ing. h.c. F. Porsche AG)

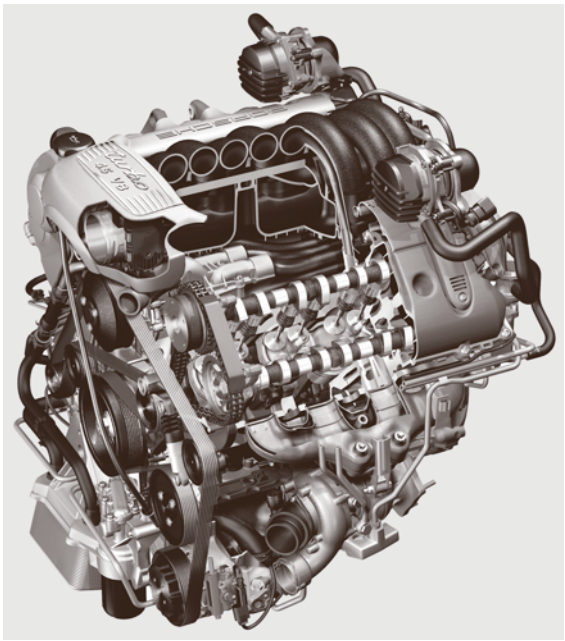


Fig. 13.53 Eight-cylinder combustion engine, turbocharged (courtesy of Dr. Ing. h.c. F. Porsche AG)

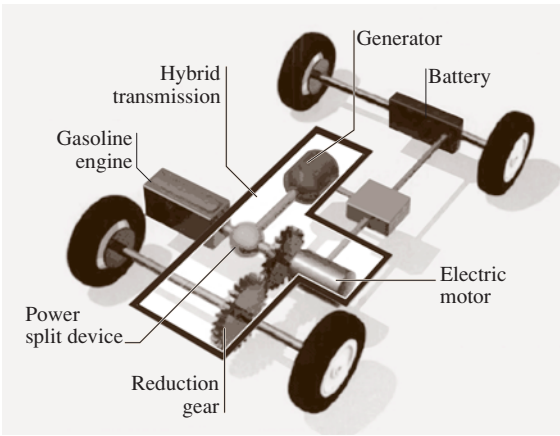


Fig. 13.54 Schematic view of a hybrid drive concept (courtesy of Toyota Motor Corp.)◀

time to time by a generator powered by a combustion engine. In parallel hybrids, the wheels are driven both by electrical motors or a combustion engine alternatively or even simultaneously. Mixed hybrids are a combination of parallel and serial power flow, and many versions are available.

The major components of hybrid drives are the combustion engine, the battery, the gearbox, and the electrical engines, which have to be harmonized for integration into the final concept.

For further information about low-emission concepts, refer to [13.47–53].

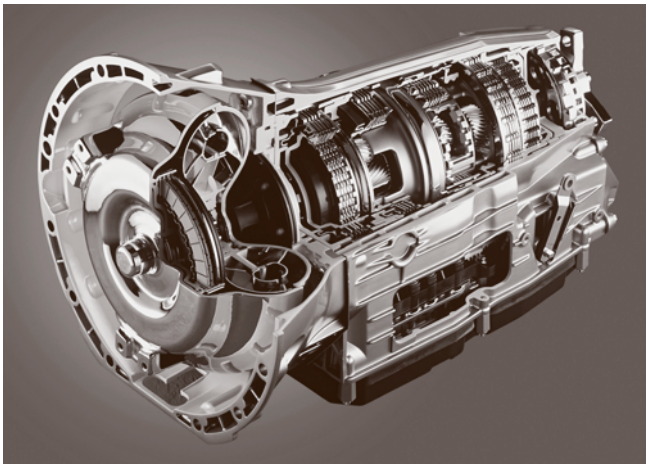


Fig. 13.55 Seven-gear automatic transmission (courtesy of Mercedes Car Group)

Transmissions. The function of the transmission is to adjust the engine speed and torque according to the needs of the current driving situation. Two basic concepts are implemented in today's cars: manual (Fig. 13.56) and automatic transmissions (Figs. 13.55, 13.57). Manual transmissions offer in most cases better efficiency compared with automatic transmissions, which in turn provide greater comfort and, depending on the experience of the driver, even optimize the utilization of the engine speed and torque in different driving situations.

Clutches. The clutch is placed between the engine and the transmission in cars with manual gearboxes, providing:

- Adaptation of engine speed during approach
- Separation of engine and transmission when shifting gears
- Safety for transmission and other components during overload conditions
- Damping vibrations

In cars with automatic transmission, the hydraulic unit of the automatic transmission provides these functionalities. For further information about transmissions and clutches, refer to [13.54–57].

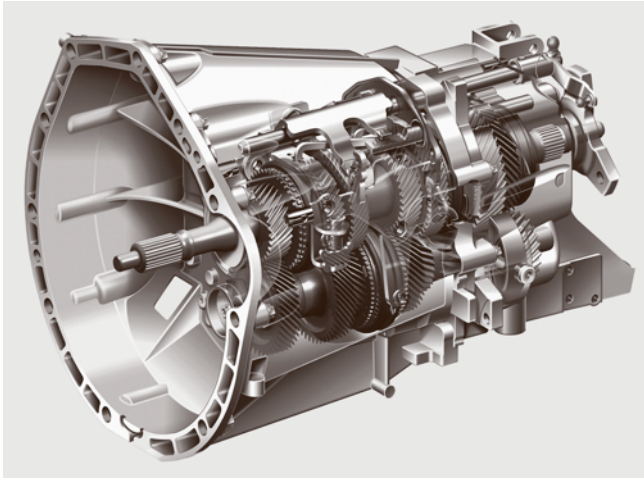


Fig. 13.56 Manual transmission (courtesy of Mercedes Car Group)◀

Interior

When designing the interior of a car, the man–machine interaction is the core issue. The interior has to be adapted to the driver and the passengers as far as possible. Ergonomics of the controls and information

instruments presented to the driver and the passengers is a major influence in order to be able to drive safely and without being stressed. Methods and tools used for ergonomic development are described in Sect. 13.2.4.

Besides the components and their placement inside the car, another important aspect of positive feeling inside the car is the interior climate [13.58]. Separately adjustable air conditioning for driver and passengers, not producing any noticeable draft, heated seats and steering wheels, as well as an air scarf for convertibles are elements which cover the needs for a pleasant interior climate in modern cars (Fig. 13.58).

Interior Materials. Individualization of luxury cars plays an important role in sales. To be able to configure a car in such a way that no-one else owns a similar one is often a very important feature for customers of these kinds of cars. Thus, providing a choice of different interior materials (aluminum, carbon, different types of wood, different types of leather, alcantara, etc. [13.59]) from which the customer can choose when configuring the car is a characteristic of luxury cars. In order to be able to provide this multitude of materials, the design of the interior components has to take account of the fact that different materials could be applied as the outer shell. Sharp edges and small radii should be avoided when planning to use leather on a surface, and the manufacturing process of placing tiny pieces of wood on the gear stick has to be taken into account when designing the standard design variant. When introducing a new, additional interior material, extensive tests (head impact, airbag functionality, etc.) have to be carried

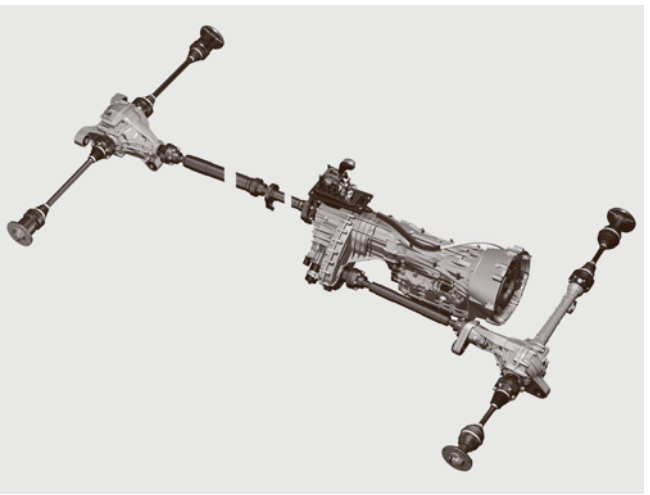


Fig. 13.57 Automatic transmission in the context of a four-wheel drive (courtesy of Dr. Ing. h.c. F. Porsche AG)

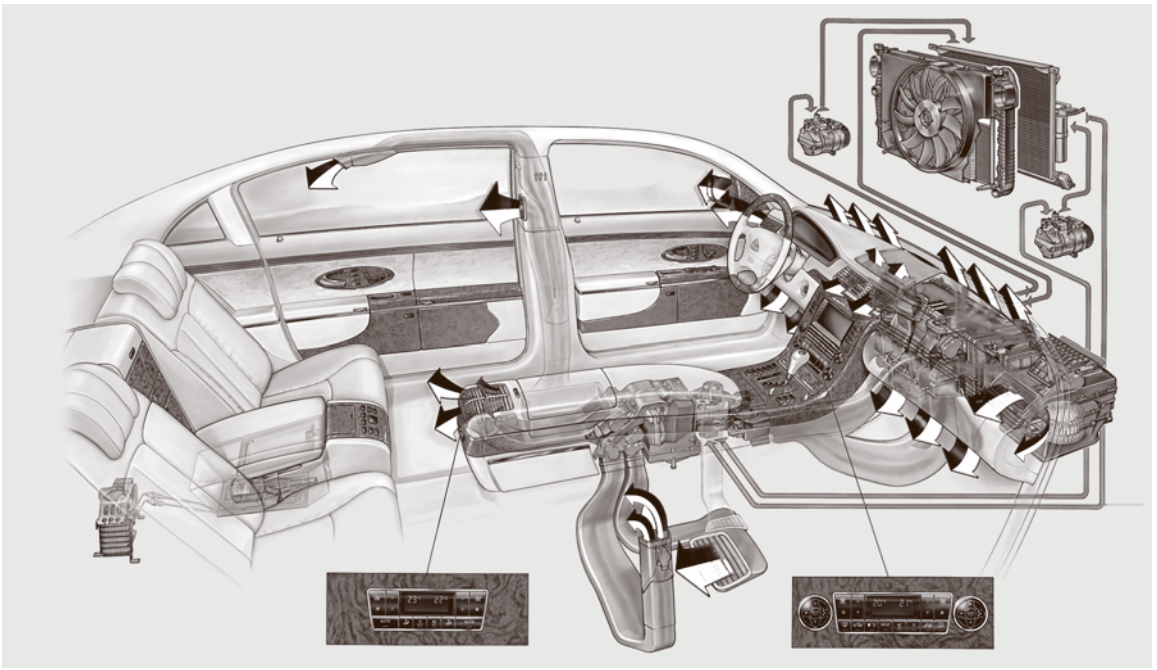


Fig. 13.58 Climate control in a luxury car (courtesy of Mercedes Car Group)

out and it has to be verified that this additional material can be handled and provided with the desired (i. e., expected) quantity throughout the supply chain.

Infotainment. Car radios are increasingly becoming integrated into infotainment components, consisting of a radio with a compact disc (CD)/digital versatile disc (DVD) changer, telephone, navigation system, onboard computer, and other information services. The benefit for the customer is that there is only one interface to deal with and that, due to the tight integration, different comfort functions can be realized more easily. An example is the integration of the traffic message channel (TMC) radio service with the navigation system, allowing for dynamic adaptation of the route in the navigation system due to TMC-based information of traffic delays.

Electric/Electronic Components

Many of the functions in modern cars cannot be realized without the intensive use of electronic components realizing, controlling, and integrating the functions into a system.

Greater demands for reliability, comfort, safety, provision of information, reduction of fuel consumption, and reduction of environmental pollution could not be satisfied by purely mechanical systems. Modern safety features such as airbags and dynamic stability programs could not operate without extensive use of electronics and microcomputers. Besides electronic hardware, software in cars is also becoming increasingly important.

Electronics can be roughly divided into:

- Sensors
- Actuators
- Controllers
- Information devices
- Network components (e.g., the controller area network, CAN)

Sensors feed information about the current status of the car and its behavior on the road into the electronic microprocessor-based infrastructure. Based on the algorithms implemented, this information is checked for necessary actions (e.g., *brake is locking – activate antilock-brake sequence*), actuators are triggered to start operation, and appropriate information for the driver and the passengers is passed forward to an information device, which may be cockpit instruments or the simple chime of a bell when seat belts are not fastened.

Figure 13.59 shows schematically the electronic elements in a car and their connections.

Sensors are intensively used in engine management, measuring, e.g., the intake manifold pressure, the mass airflow, the temperature at different locations in and around the engine, the engine speed, the crankshaft reference position, enabling correct triggering of the pulse for injection cycle, the throttle position, the knock sensor, allowing lower-quality fuel to be used as a fuel, and the lambda sensor for catalytic operation. For driving dynamics and safety purposes, sensors measure various parameters from chassis, including wheel speed – also relative wheel speed between different wheels – steering-wheel angle, and steering shaft torque, the chassis acceleration or deceleration, and the brake system pressure.

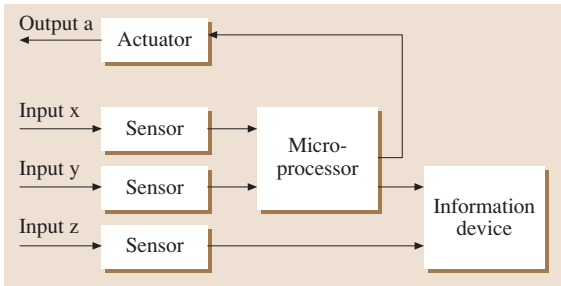


Fig. 13.59 Electronic elements in a car and their connections

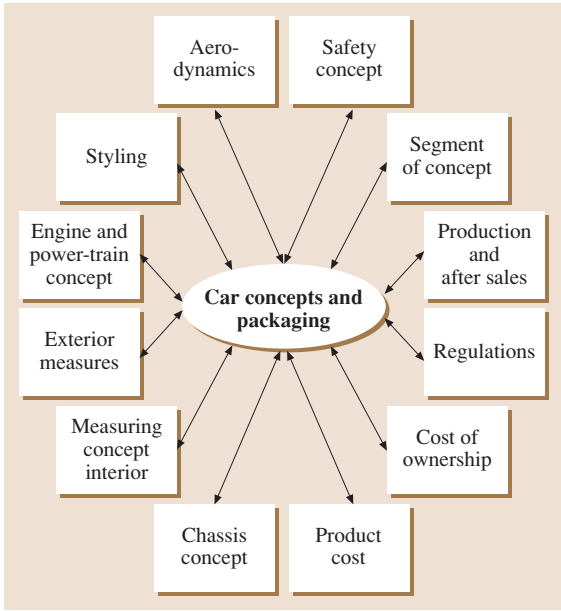


Fig. 13.60 Requirements to be taken into account when defining a car concept [13.15]

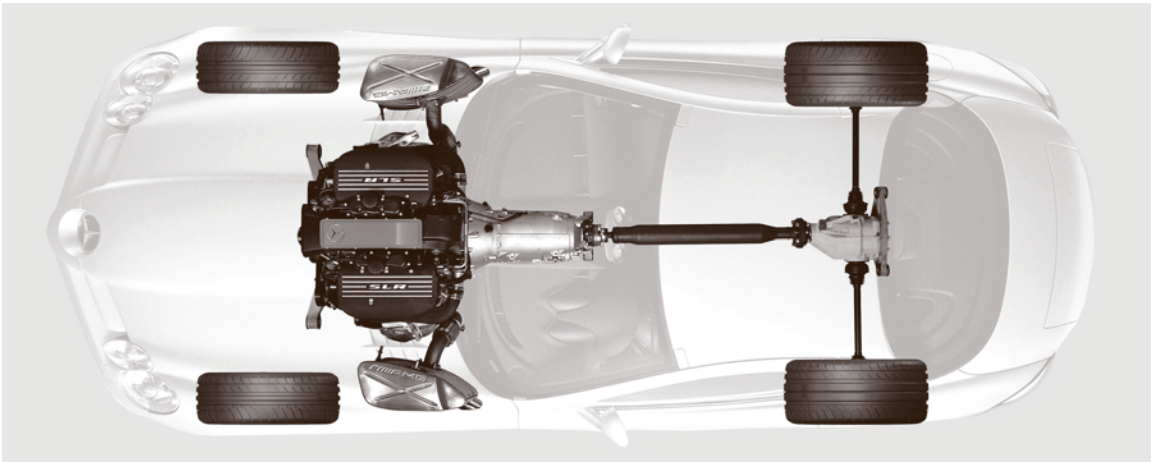


Fig. 13.61 Integration of engine and power-train components into the car concept (courtesy of Mercedes Car Group)

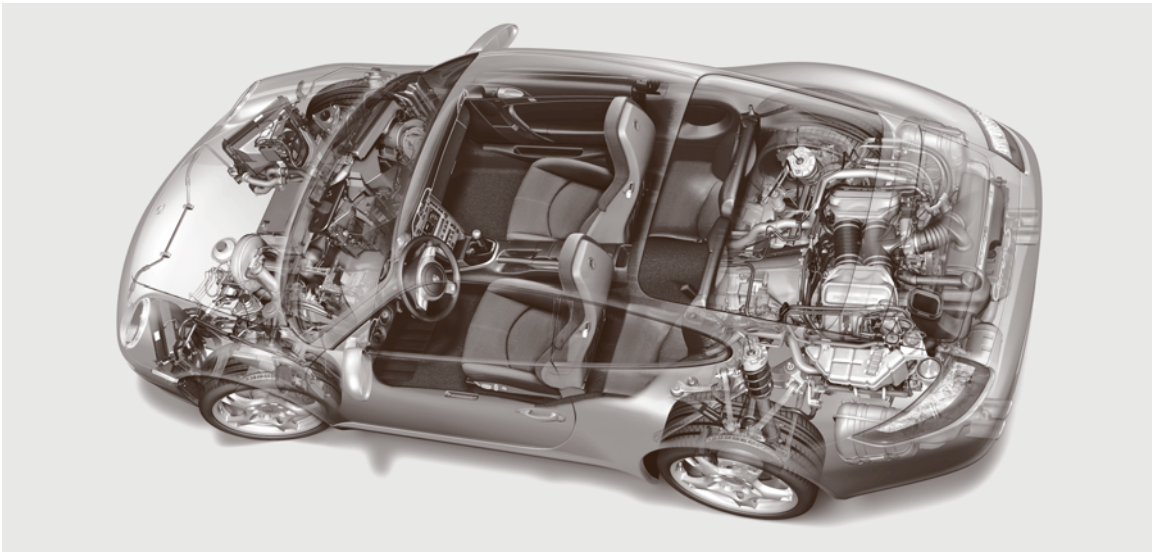


Fig. 13.62 Packaging of components in a sports car (courtesy of Dr. Ing. h.c. F. Porsche AG)

Due to their physical working principle, sensors can be divided into various categories (Table 13.5).

Besides sensors, actuators are another important group of elements in an electronically operating system. Actuators range from starting generators, supplementary drives to different actuators in the surrounding of the engine (crankshaft actuation, electromechanical gear shifting), those used for implementation of safety features (tightening seat belts, electrical brakes, electrically actuated rollover safety in convertibles), for comfort functions (powered windshields, powered sunroof, memory-based electric seat readjustment), for information functions (CD player, navigation system), and to provide active light management depending on the external brightness, to mention only a few.

With the large number of electric and electronic components in a car, electrical energy consumption becomes a critical issue. Even during standstill, most modern cars need a certain voltage to be in a standby position for remote opening systems and for theft-protection devices. Since especially in cold outside temperatures, the driver and the passengers demand a large number of electrically supported functions (defrosting, seat heating, etc.) the management of these

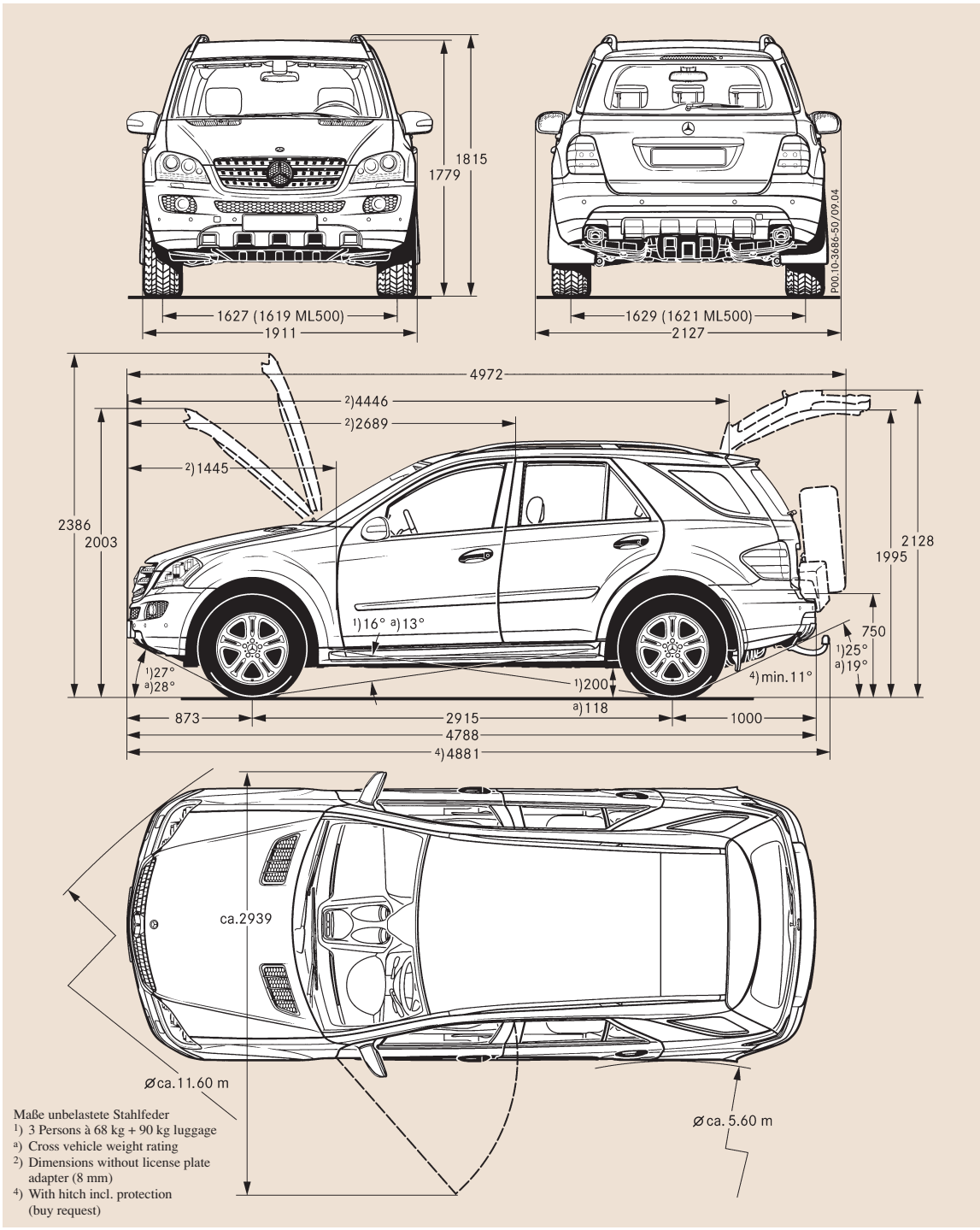


Fig. 13.63 Dimensional concept of a sports utility vehicle (courtesy of Mercedes Car Group)

Table 13.5 Overview of sensor categories [13.15]

Magnetic sensors	Revolutions, position, angle
Electrical sensors	Position, angle, seat occupation identification
Micromechanical sensors	Knock, pressure, acceleration, torque
Chemical sensors	Air quality, composition of exhaust gas
Thermal sensors	Intake air mass flow, part temperature
Ultrasonic sensors	Distance, relative speed
Radar sensors	Distance, car surrounding
Optical sensors	Position identification, driver identification, object identification

units in order to cover the demands of the passengers while enabling the car to get into a stable operating condition is a challenging task. Reaching stable (i.e., constant) operating conditions in the electric network of the car can sometimes take over 1 h. Further increases in the need for electrical power in cars is leading to the development of 42 V electric systems instead of today's 12 V electric system [13.60].

Nearly every area of car functionality is supported by or implemented in combined electronic hardware–software systems. The most challenging of these systems are these for safety: clear determination of a situation where airbag deployment is needed, and airbag deployment at exactly the right time in the right fashion is a highly complicated systems design task. Precrash sensing and initiating preventive actions such as tightening the safety belts, closing the sunroof, etc., which are even more demanding functions than crash detection, are now available first in luxury cars.

For further information about electrics and electronics in automotive application refer to [13.62–69].

Packaging

Packaging's primary task is to collect the different requirements and targets for a car and implement them

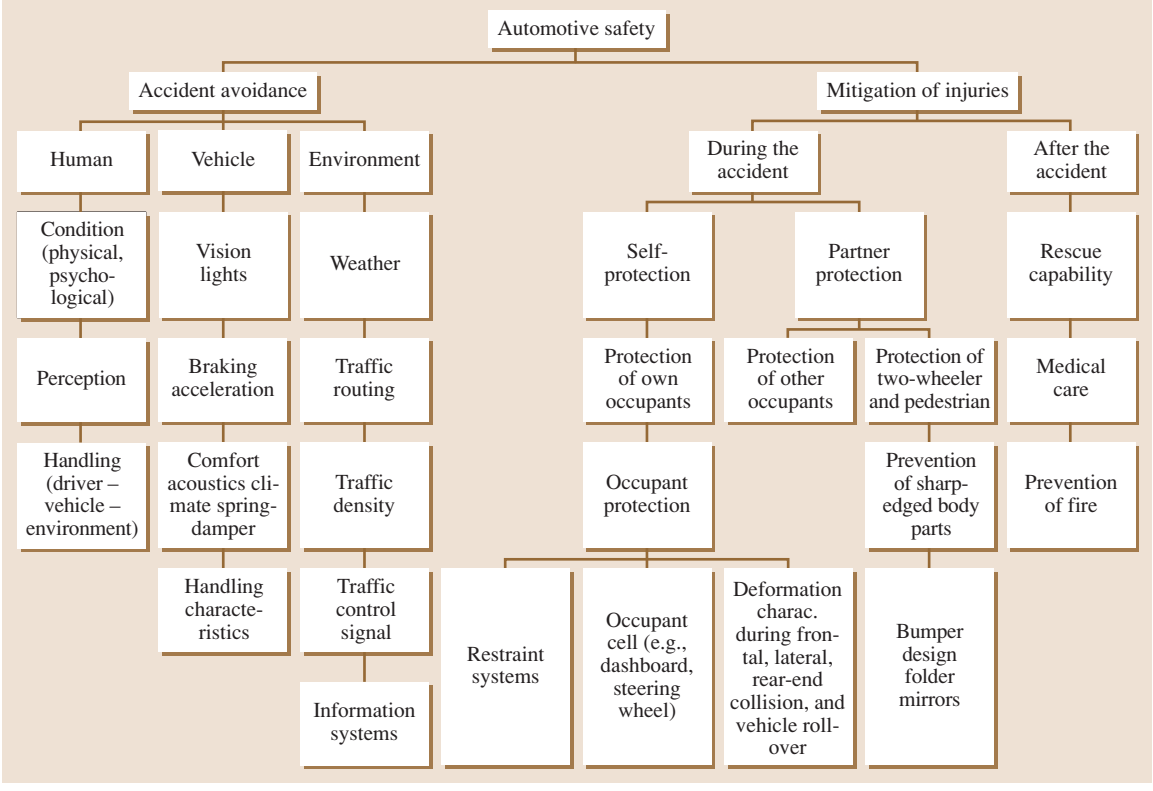


Fig. 13.64 The field of automotive safety [13.61]

Fig. 13.65 Crash simulation and hardware crash (courtesy of Mercedes Car Group)

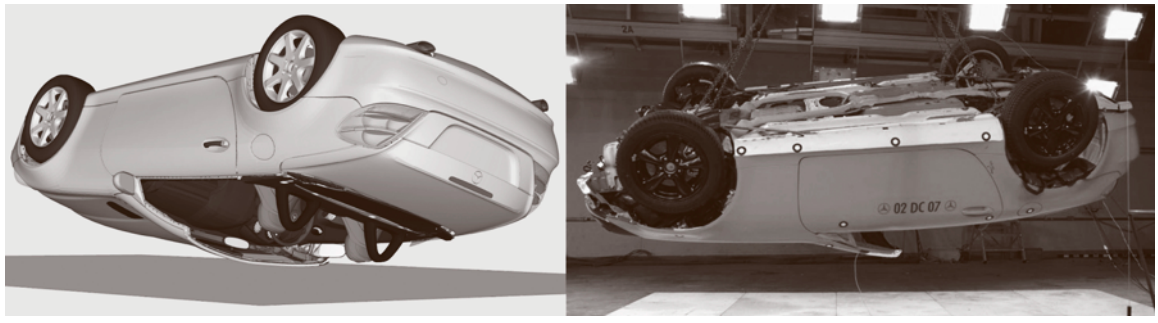


Fig. 13.66 Crash simulation and hardware crash of a convertible (courtesy of Mercedes Car Group)

into a basic design with main dimensions and definition of the placement of the most important components in the car (Figs. 13.60–13.63).

The result of packaging is a compliant arrangement of components in terms of space and functional dependencies. The major influences on the defined car concept resulting in the car packaging are [13.15]:

- Technical data
- Primary competitors
- Range of use (leisure, city, off-road, sports, etc.)
- Body versions
- Safety concepts including crash structures
- Seating capacity and ergonomic requirements for seats
- Trunk volume
- Engine and power train concepts

One of the results of packaging is the dimensional concept of a car where all main measures of the car are defined.

For further information about packaging and ergonomic aspects in automotive engineering refer to [13.70–72].

Automotive Safety

As well as dealing with crash safety, automotive safety is a broad area covering many aspects. Figure 13.64 gives an overview over the whole field of automotive safety.

Active and passive safety are prime selling arguments. The responsibility for the health of the driver, the passengers, and other traffic participants is also taken very seriously by automotive manufacturers.

Active safety is the area of measures aiming at avoiding an accident. Passive safety is the area of measures minimizing the consequences of an accident when it happens. Besides technological provisions in the car concerning safety, the factor most strongly influencing the probability of an accident and its consequences is the driver: careful, defensive driving is the best precaution.

Various coefficients describe the criticality of body stresses induced by an accident, and regulations demand that test data are below a certain level for each of these. These coefficients include the head protection criterion (HPC), the thorax compression criterion (TCC), and the tibia compression force criterion (TCFC).

Table 13.6 Dummy types defined in *Federal Motor Vehicle Safety Standard 572: Antropomorphic Test Devices* (http://www.nhtsa.gov), NHTSA, Washington USA

FMVSS 572 Subpart B	50th percentile male
FMVSS 572 Subpart C	Three-year-old child
FMVSS 572 Subpart D	Six-month-old infant
FMVSS 572 Subpart E	Hybrid III test dummy
FMVSS 572 Subpart F	Side impact dummy 50th percentile male
FMVSS 572 Subpart I	Six-year-old child
FMVSS 572 Subpart J	Nine-month-old child
FMVSS 572 Subpart K	Newborn infant
FMVSS 572 Subpart L	Free motion headform
FMVSS 572 Subpart M	Side impact hybrid dummy 50th percentile male
FMVSS 572 Subpart N	Six-year-old child test dummy, beta version
FMVSS 572 Subpart O	Hybrid III 5th percentile female test dummy, alpha version
FMVSS 572 Subpart P	Hybrid III three-year-old child crash test dummy, alpha version
FMVSS 572 Subpart R	CRABI 12-month-old infant crash test dummy, alpha version

Designing the car to meet these regulatory demands impacts on the car structure, the materials used, and the shape and location of interior components, and safety measures such as airbags and their deployment algorithms have to be adapted.

For the verification and validation of the crash performance of a car, crash simulation and hardware crashes are used (Figs. 13.65 and 13.66).

For the tests, various dummies are used. Table 13.6 shows the different dummy types used in today's crash tests.

For further information about automotive safety refer to [13.73–76].

13.2.3 Car Development Processes

Overview

The development of cars is – like every other product development – a company-specific process. Many influences such as the size and level of globalization of the company, its product portfolio, the number of produced cars, and the proportion of outsourcing during the product generation process affect the characteristic of the specific development process.

Generally, the development of cars consists of three main elements:

- The strategic phase
- The car development project
- The adaptation phase

The following picture gives an overview of the car development process (CDP). Like any other product development process, car development starts from the global point of view and the rough concept.

During the strategic phase the car is planned, one could say, from above. Aspects considered at this stage come from the environment, the market, and the company itself and lead to the strategic specification of the car.

Aspects that influence strategic decisions are, for example, social trends such as increasing awareness of environment or the increasing necessity for safety, the need to substitute technologies as a consequence of improvements or to provide replacing technologies as a political move (e.g., the fuel cell as an alternative drive mechanism), the product portfolio of competitors, rationalization of the companies own workflow, and new functions to be implemented in order to make the product more attractive such as active light illumination or headway distance control.

The variety of aspects influencing the strategic definition of the car is huge but there are just as many approved methods to support this phase of economical and political decisions. The methods are general and there are many good summaries on this topic in the literature [13.77].

The main specifications which have to be defined during the strategic phase of a car are summarized in Fig. 13.67. One important boundary condition for the development of a car is the definition of the product family. For this step the needs and trends of the market have to be analyzed, the product portfolio of the competitors compared with the company's own portfolio, and the profitability of the possibilities calculated.

By defining the key performance indicators the characteristics of a car are determined which are decided by the customer target group which will be reached. Besides the factors price, design and general affection for the brand have a big influence to. Examples are the horse power to be supplied, the noise and

consumption allowances, the kind of air-conditioning system, and the need for automatic locking of the entire car. The basis for the definition of the key performance indicators and the major technical features to be implemented are again the general trends and circumstances in the market. Moreover, methods and technology are advanced in principle in order to achieve competitive advantage. The paragraph *Project-Independent Predevelopment* focuses on car-specific development in this context.

If an advantage is to be expected when, for example, an important customer need will be fulfilled (e.g., providing driving comfort in what is actually a sporty car by an additional feature called active body control) or requirements resulting from political discussions can be satisfied (e.g., discussions of the wrecked-car regulation) then advanced development will be described. The achievement aimed for in this context can be the result of project-independent or project-specific predevelopment, but is also a topic of general research. In this case the results are not available at the time of description. During the verification of the concept, which is an early stage of the car development project, it has to be established whether the research is ripe to be realized or not. In the paragraph *Project-Specific Predevelopment*, the aim of this kind of development will be made transparent and examples will be described.

A further fundamental declaration at this stage of car development is that the project promises to be profitable. A car can be profitable when it achieves a monetary profit due to the price and the calculated number of pieces planned. On the other hand a car can also be nonmonetarily profitable, as is the case when it is of strategic importance (for example, the Maybach or SLR for the Mercedes Car Group, of the Phaeton for VW).

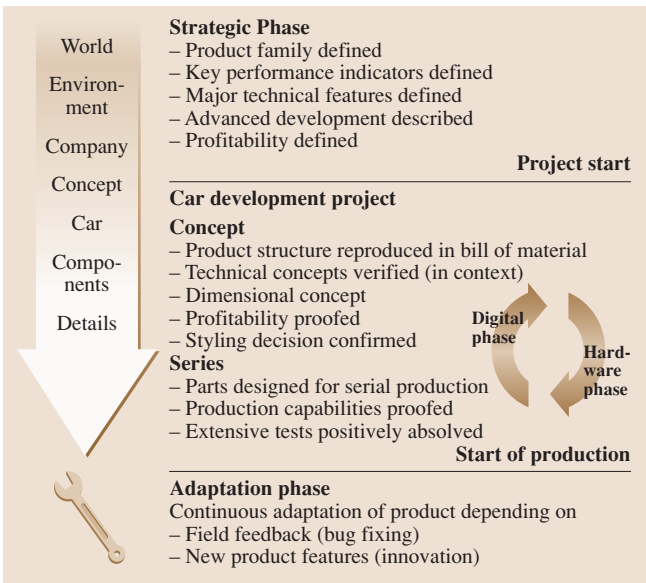


Fig. 13.67 Car development process

The car development project describes the period of time during which the car and its production is specified in all necessary details in order to start the so-called *job number one*, meaning manufacturing.

The car development project (CDP) is carried out in two steps: development of concept and series (Fig. 13.67).

In the concept phase it is determined whether the set of rough outlines and technical features defined during the strategic phase are compatible. For this purpose the dependent system or even the entire car has to be investigated in context. By defining the dimensions of the concept the size of the car is determined and with it one major specification of the entire car set. Already at this stage of development all outer measurements are fixed and the functional spaces inside the car, such as overhead space, leg room or trunk capacity, are provided (Fig. 13.68).

Parts and components are defined and created. The basis for car development however is the product structure of the car, which is reproduced in the bill of materials (BOM). The bill of materials is the connection to production, regardless of whether it is a prototype or the series model that is to be manufactured. Since the configuration of the car and its possible variants are realized within the BOM, the BOM is used to order materials, parts, and components and hence forms the basis for logistical processes. In the section *Product Structure and Bill of Materials* the importance of the product structure with respect to the BOM is clarified.

Furthermore, styling decisions generally have to be confirmed during the concept phase. One has to ensure that the functional components can be accommodated and that the dimensional concept can be realized within the body prepared by the styling. Although changes of design will be accepted during the concept development, the fundamental design features are agreed upon before the car project starts.

Two main streams of development overlap in the development of both the concept and the series vehicles: the digital and hardware phases (Fig. 13.69).

Development always starts with the digital phase. Parts and components are designed for prototype or serial production in computer aided design (CAD) systems. Assemblies are generated using the bill of

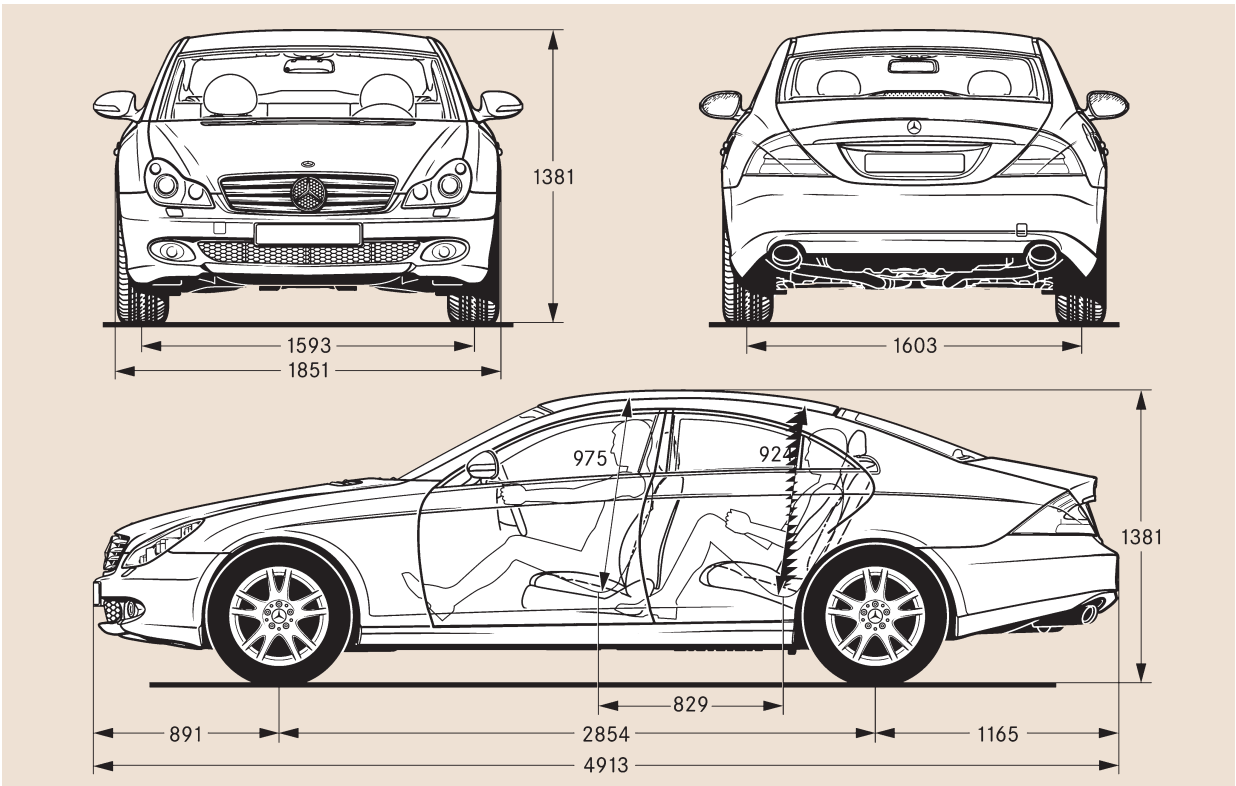


Fig. 13.68 Dimensional concept of Mercedes CLS (courtesy MCG)

materials, product data management (PDM), and CAD functionality (see the section of *Product Structure and Bill of Materials*).

The digital model of the car is used for simulations such as package investigations via a digital mock-up (DMU) or kinematics calculations to ensure stability or check thermal behavior, as well as simulations to plan the production process.

After a certain delay the hardware phase starts. Here the digital models are physically generated (in the form of a physical mock-up) in order to verify the digital simulations. The tests conducted vary depending on the level of progress of development. Examples are tests of concept and structure, functional tests of components on a mule or in a running car, and crash tests to ensure the safety features of the car in different accident situations and for different passenger combinations. The results of these tests are used to optimize the simulation models for the digital testing (for more details on interactions between the digital and hardware phases see also Sect. 13.2.4).

When the concept is verified, the series development starts, which means that parts and components are now redesigned for serial production using the results from tests carried out during the concept phase. Extensive tests are used to try to ensure the quality of the mature product. Whether specific noise requirements are met and whether the car works under extreme climatic conditions can only be fully verified by testing the car in its final realization. In preparation for production the capabilities of the plants also have to be proofed.

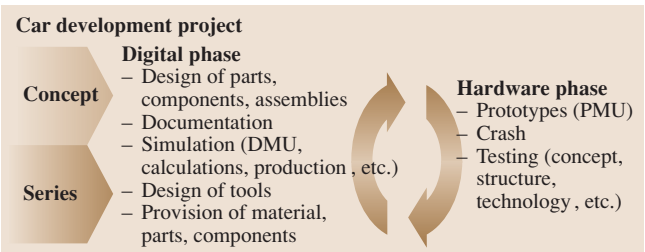


Fig. 13.69 Car development project (CDP)

After a specific level of development is reached, the tools to produce the parts will be designed, and materials, parts, and components delivered by suppliers are ordered on the basis of the digital models, organized in the product structure of the BOM.

At a certain point, the start of production (SOP), the development has to be completed and the car will be produced. The development however is not yet finished, because continuous adaptations of the product are necessary due to field feedback (bug fixing) and innovations that should be implemented in order to stay competetive.

Project-Independent Predevelopment

Aside from car development projects intending to result in the series production of cars for end consumers, there are also development activities as a preparation to be able to do so.

These activities can generally be divided into different classes (Fig. 13.70).

The main distinction in project-independent predevelopment is between methods, technology, and styling concepts.

The definition of key technologies and features where there is substantial differentiation from other OEMs is a strategic trigger for the definition of predevelopment projects.

Roadmaps of upcoming regulations in the different markets are also used to derive predevelopment projects.

Methods. The demand for shortening product development time, reducing cost and development risk, providing new features (e.g., direct fuel injection), and coverage of new regulations (e.g., pedestrian protection) lead to the need to develop new methods to cover these demands. Thus, the methods can also be divided into methods for improving efficiency and those for meeting new demands.

All of these methods are usually developed with a specific target car development project in mind, where they will be used the first time. Evaluation of these methods is either a separate project or is done during the real application of the methods in a car development project, in many cases accompanied by conventional fall-back or back-up methods if the new method does not work properly. An example is the use of computational fluid dynamics (CFD) methods for aerodynamic layout, where the fall-back method is a scaled model in a wind tunnel.

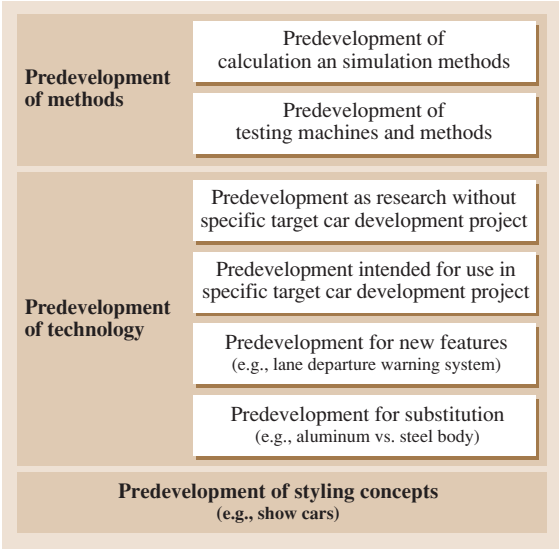


Fig. 13.70 Different kinds of predevelopment projects

However, general research done outside the OEMs also influences car development; for example, developments of new materials (alcantara or plastics), new manufacturing techniques that achieve greater stability (tailored blanks), and mathematic algorithms for calculating stress and strength in parts such as finite element modeling (FEM) are successfully applied in car development today.

Technologies. Technologies are predeveloped, driven either by external demands such as regulations or by the company's own impulse. One internal trigger may be the need to reduce weight in a certain area of the car in order to meet future weight and weight distribution targets for new car concepts, thus leading to the development of substitute materials, part structures, connection technologies, and production technologies. Another trigger is the fulfilment of expected customer needs for comfort (automatically shading roof glass), safety (lane departure warning systems), and entertainment (several infotainment components) [13.78, 79].

Styling Concepts. Styling concepts are developed

- In order to check whether a certain styling proportion or styling theme will meet the customer taste, especially for the intended market segment
- To identify whether there is customer acceptance for a completely new car (and thus styling) concept

- To show the fitness of the automotive company, and for marketing purposes

Styling concepts that receive a positive resonance from the customer base are often used for car development projects. Here it is especially important to reach the SOP in a short period due to shifting customer tastes. Car shows such as those in Detroit or Geneva are often used to present these styling concepts to a broad group of people and to obtain direct feedback or feedback through the press.

Dependencies. One example of the need to develop new methods and technologies in the past was regulation for pedestrian protection, which is gradually being implemented in different stages. To cover the demands of pedestrian protection, various predevelopments had to be undertaken, including:

- Methods for simulation of the body under contact with the car front and the resulting load on the car body parts
- Methods for calculating the critical points in the car exterior which have to be analyzed
- Methods for testing a real car front with a dummy head and measuring the deceleration of the head
- Car concepts and technologies to meet the demand for deformation zones in the affected area of the car front (passive technologies such as providing enough space between the exterior shell and the underlying stiff components, and active technologies such as parts of the car front being raised during impact to enlarge the deformation zone and distance)
- Styling concepts for different variants of car fronts taking passive technology with larger deformation zones into account

Project-Specific Predevelopment

Unlike project-independent predevelopment, activities in project-specific predevelopment focus on the approval for the start of a car development project.

In project-specific predevelopment various activities take place, including:

- The packaging concept for the car is roughly defined.
- Technologies to be used in the car – also from project-independent predevelopment – are identified and decided on.
- Styling proportions and major styling elements are decided on.
- Targets for cost and weight are defined in the upper levels of the product structure.
- Strategic development partners for specific components and/or modules are selected.
- The time frame for car development until SOP and major milestones are defined.
- The project organization is defined at least at the upper levels; the need for engineering resources is roughly estimated.
- The expected economics of the project over its lifetime until end of production and spare parts provision is calculated based on extrapolation from previous car projects.

If the technological and styling concepts meet the expectations of the management, the economics is expected to be above the margin defined in the automotive company, and if the project can be handled with the existing resources, the project is likely to be approved.

Concept Development

Concept development starts when the car development project starts, and ends when series development starts, usually with the milestone of *styling freeze*, at which point styling hands over leadership to body-in-white (BiW) development. Its main goal is to define the technological concepts and secure their function in the context of the car.

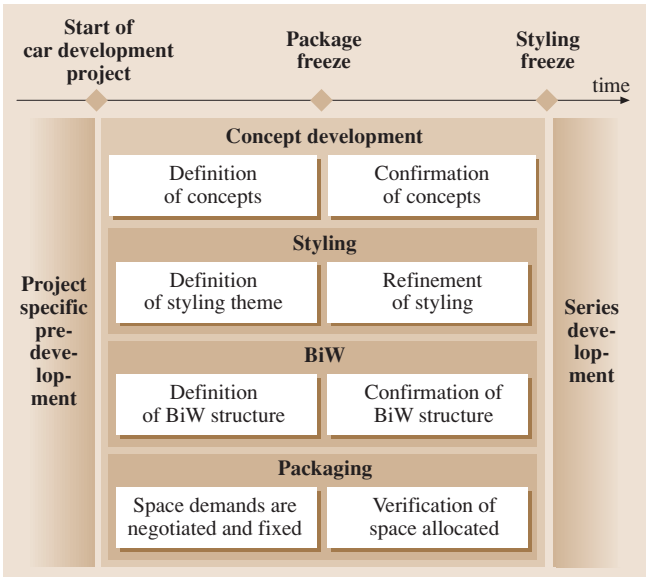


Fig. 13.71 Overview of activities of concept development

At the start of concept development, many activities start in parallel. Exterior and interior styling, packaging and ergonomics, body-in-white design, aero-/thermodynamics, and concepts of all other technology departments are defined and validated in the functional context of other concepts and taking space restrictions into account.

In Fig. 13.71, the main activities of concept development are shown at an abstract level.

Package freeze is a milestone in concept development when all space demands from all parties have been announced, and conflicts have been resolved by trade-offs under guidance from packaging. From this point onwards, violations of space utilization are monitored by packaging and measures for resolving the conflicts identified.

Near the end of concept development, the first hardware prototype is crashed to confirm the structural concept of the car.

During concept development the set of options from which the customer will be able to choose when ordering a car is confirmed. This set of options will be developed during the car development project as well, bearing in mind that there will be different SOPs for the different options after the SOP of the car with the first set of options to be offered. During car development there will be further decisions relating to additional options, resulting from market research and customer feedback, which will be integrated into the project as well (Fig. 13.72).

Series Development

Depending on the internal definitions of the automotive company, series development starts with the milestone of styling freeze or styling release and ends at the start of series production (SOP). The major task during series development is to transfer a car concept that has been proved during concept development to a car series that can be manufactured under series conditions in high volume, at high quality, in the variants offered to the customer, at a competitive production cost.

For this purpose, a number of milestones have to be passed:

- Prototype tests are finished with positive results.
- Tools for series production are released.
- Contracts with the suppliers for series production are concluded.
- All regulations and country-specific requirements are fulfilled.

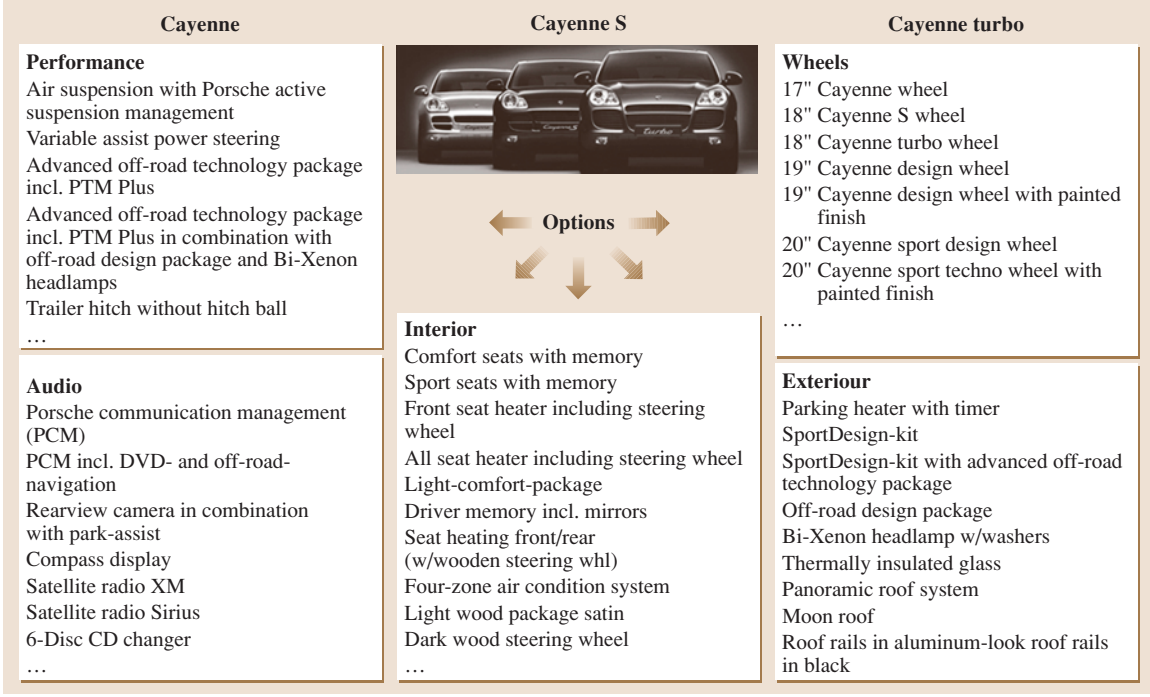


Fig. 13.72 Options the customer can choose from – example of the Porsche Cayenne

- Sample delivery of parts produced with series tools are positively approved.
- Logistics processes are defined and verified; ability to handle the logistics according to the planned production volume is proved.
- Marketing concept (advertising strategy) is defined.
- The dealer organization is informed about the product and its maintenance.

Series development is done in even closer cooperation with the suppliers than concept development, since sourcing for all parts is defined during series development, whereas during concept development often only strategic suppliers are directly integrated in the car development process.

With the start of series development at the latest, the project team structure is defined and fully operational. The teams consist of members of design, production, sales, purchasing and quality and are responsible for the development of the assigned components within the given limits of cost, time, quality, functionality, and performance.

Activities After Start of Production

After the start of production, the car development project is finished. At this point two major activities are still ongoing:

- Development of further options to be offered to the customers at a later point of time
- Support of series production

Further options are usually planned with the car development project in total. The production schedule of these options is integrated into the overall production scenario of how many cars will be sold at which time in what markets.

Support of series production is an ongoing activity which is usually not planned together with the car development project but is rather a support function which is continuously improved. Here:

- Changes of parts are made in order to optimize production processes in terms of cost and reliability.
- Design changes according to feedback from the customers and dealers are implemented.
- Design is optimized in order to reduce material cost.

Depending on the type of change of the part design, effects have to be taken into account up to spare parts provision, thus each change is evaluated in detail in terms of overall cost and benefit.

Depending on the product life cycle and competitors' product developments, facelift projects are established in order to upgrade a car which has been in the marketplace for a certain time with additional features, better performance, and adapted styling. After another period of time, the next generation of car development starts.

13.2.4 Methods for Car Development

As with the development of most technical artifacts, car development makes use of various methods for defining and verifying form, function, and performance. The methods used can be categorized into:

- Virtual methods
- Hardware methods

Virtual methods operate on a product description, usually contained in a computer-based product definition. Computer-aided tools are used to generate and verify the current state of the product definition (Fig. 13.73).

Hardware methods make use of real parts, components, assemblies or products (in car development projects, usually prototypes and pre-series cars). On testing machines, on special driving courses, or out on the street, the product or parts of it are tested for function, performance, and durability (Fig. 13.73).

Virtual and hardware methods usually go hand in hand when a car is being developed (Fig. 13.74).

Generally, with virtual methods a large number of design alternatives can be generated and/or evaluated in a comparatively short amount of time, whereas hardware methods cover more aspects of the real behavior of the product, since not all aspects of product behavior are modeled in the virtual methods.

In the following some of the most important methods for car development will be highlighted and it will be explained how they are used in the development process.

Methods for Product Layout and Conceptual Development

In the early phase of a car development project, styling, package, the body-in-white structure, and aero- and thermodynamics are the key factors which have to be harmonized in order to generate a suitable overall car concept.

Styling

Virtual. Computer-aided styling (CAS) tools allow the definition and manipulation of two-dimensional

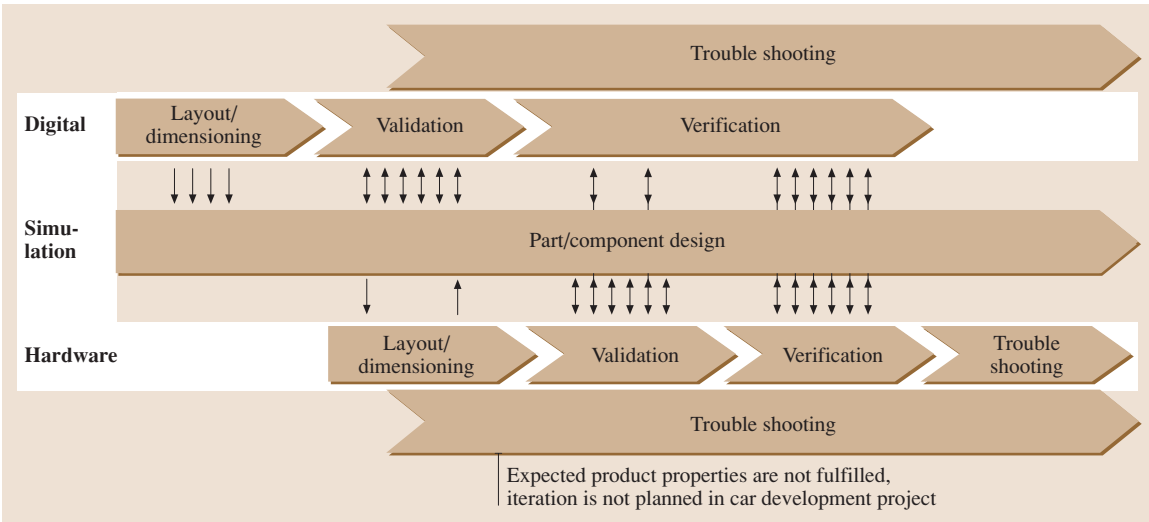


Fig. 13.73 Support of part and component design by virtual and hardware methods

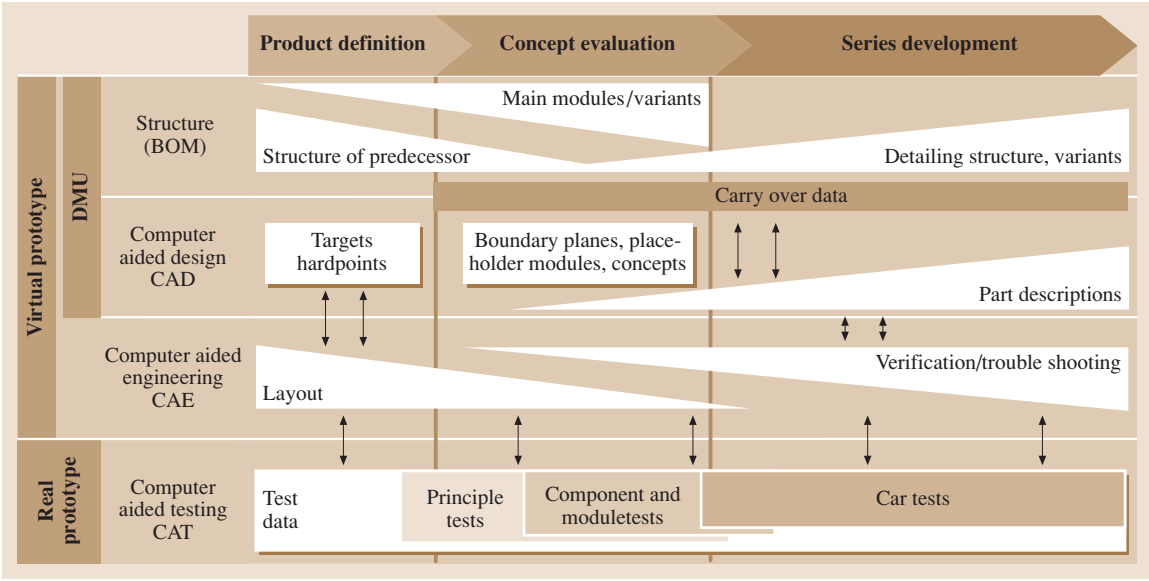


Fig. 13.74 Milestones of virtual development and verification in hardware [13.80]

(2-D) sketches and three-dimensional (3-D) geometry, and have special capabilities for texturing and shading in order to obtain the most realistic pictures or models from the styling definition. Visualization can be done using high-definition printers or monitors, and for realistic 3-D-viewing virtual-reality (VR) equipment can be used. CAS tools allow for the conversion of styling data directly into CAD data, thus making it possible to shorten the time between finishing a styling model and checking, for example, that the car package complies with it. The data transferred into more technically oriented CAD programs can be used as a base to design metal sheets for the body shell. This is a method to define the geometry needed for finally milling the stamps for the outer shell parts, for example. Also, the airflow around and through the car can be simulated using these models.

The major shortcoming of virtual styling process today is that evaluation of a virtual styling model is far more difficult than that of a hardware styling model. Even most sophisticated VR applications and equipment still do not give the same impression as a real car model in real-world surroundings. Therefore, hardware styling models are essential in the styling process as well [13.81].

Hardware. In the early phases of the styling process, scale models (1 : 4 to 1 : 3) can be used to show different proportions and styling topics of a car to be developed. One (or more) of these will be decided on and then refined and used as the base for further styling activities.

Tape drawing is a method for defining the main contours of a car or areas of a car at a scale of 1 : 1 as 2-D models. Black tape is glued to a white surface for the development of styling concepts and their fast visualization and adaptation.

Styling models at a scale of 1 : 1 are mostly made out of clay, and then finished with painted foil to give the styling model a more realistic appearance. Evaluation of the styling model should be done under different light conditions in different surroundings. With special preparation, the styling model can also be used for aerodynamic evaluation. Changes necessary in the hardware styling model to address aerodynamic needs are less easily implemented compared with the adaptability of a virtual model.

Interaction Between Virtual and Hardware Models. Virtual styling models can be milled in styling foam, a painted foil can be applied, and the model can then be used for evaluation purposes. The milled model can also be worked on further with conventional methods (clay modeling, Fig. 13.75), and later the modified model can be digitized with 3-D digitizing machines (3-D scanning).

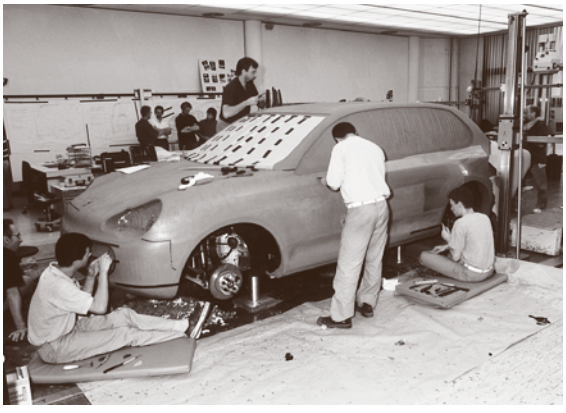


Fig. 13.75 1 : 1 clay styling model (courtesy of Dr. Ing. h.c. F. Porsche AG)

For interaction with other processes in car development (technical checks on the car package, straken) it is important to ensure that the version of the styling model evaluated is the same version that is used for technical checks and strak activities. For this purpose, team data management systems are used.

Both virtual and hardware models are used (Figs. 13.76, 13.77):

- For purposes internal to styling (evaluation, discussion)
- For evaluation by management
- For checking the acceptance of the styling by potential customers (car clinics)

During these car clinics, one or more alternative styling models intended for use in a car development project are usually presented, together with models of the main competitors' models in the target market segment. Customers' responses are used to validate whether the styling fits the taste of the customers, and that the styling is sufficiently differentiated from that of the competitors' cars.

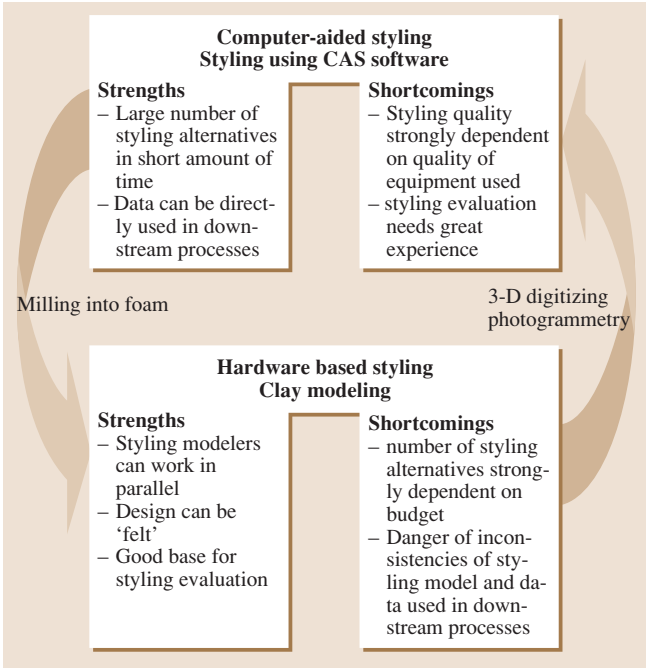


Fig. 13.76 Interaction of virtual and hardware methods in styling

Table 13.7 Strengths and weaknesses of various methods and tools for ergonomics development [13.71]

Method/tool	Strengths	Weaknesses
Design recommendations and checklists	• Quick • Easy to use	• Relevance to specific users, tasks or vehicle type may be dubious • May have little scientific validity • No account taken of compromises • Either too specific (i. e., should be 457.2 mm) or too general (i. e., should be comfortable)
Anthropometry and biomechanics	• Quick • Good for novel designs • Useful for asessing the influence of age, sex, race, etc. upon design	• May be a lack of data relevant to user or task • Data may be out of date • Data often relate to standardized postures, not necessarily working postures • Design may become too academic, mistakes being hard to identify
3-D human modelling CAD systems	• User- and task-specific predictions, quick and accurate for geometric issues such as fit, reach, and vision • Enables effective communication at an early stage • Compromises can be objectively explored	• Expensive to set up (hardware, software, training), but very cost effective thereafter • Does not assess personal preferences, psychological space, fatigue, task performance
Mock-ups and fitting trials	• Control selection of users and their tasks • Study comfort and performance over time • Sound basis for identifying good and poor designs using both objective and subjective methods • Essential for novel designs • Compromises can be investigated • Design problems are quickly identified	• Can be time consuming and expensive • Can be difficult to obtain representative subjects • May not be a very realistic simulation of task or environment
Owner questionnaires and interviews	• Valuable information direct from the user population • Small details may be detected which the casual observer may have overlooked • User involvement	• User may take poor design for granted • Opinions can be strongly biased • Cannot be used for novel designs until after production • Biased sample – does not include those people who chose not to use the existing equipment • Biased sample – low response rate from postal questionnaires, who returns them? • No detailed assessment of body size, performance or comfort
User trials and road trials	• Control selection of users and their tasks • Study comfort and performance over time • Sound basis for identifying good and poor designs using both objective and subjective methods • Allows comparative testing	• Can be time consuming • Require production and/or prototype vehicles to test • Can be difficult to obtain representative subjects

Packaging and Ergonomics

Virtual. Packaging in the early phase of a car development process means defining prescriptions for space utilization by the different technical departments. Also, the main space features of a car such as the size of the trunk are defined here. Space prescription is usu-

Table 13.8 Sample chart for definition of car configurations for geometrical space checks using DMU

Geometry influencing specifications	Car configuration 1	2	3	4	5
Left-hand drive	×		×		×
Right-hand drive		×		×	
Automatic transmission	×	×			×
Manual transmission			×	×	
Small engine	×			×	
Large engine		×	×		×
Convertible			×		×
Coupé	×	×		×	

ally done using packaging boundary shapes which are used to define the most critical areas in terms of building space in a car. They also reflect geometrical demands derived from regulations such as minimum sight angle etc. Ergonomics of a car (driver position, accessibility of steering wheel and switches, driver's view) are checked by using virtual manikins, representations of human bodies which are derived from a statistical analysis of the main geometrical body features of human beings. One example is the manikin RAMSIS [13.82], which is a software package that can be used for any kind of geometrical simulation of a human being, and can be integrated into CAD models (Figs. 13.78, 13.79).

Various methods and tools are used in car development to identify the needs of the customers and thereby to derive design guidelines for the design, shape, and location of interior components. Table 13.7 gives an overview of methods for ergonomics development.

Packaging in the early phase is also responsible for the interface between the *technical world* and the *styling world* and trade-offs between them.

During the car development project, the packaging activities change from prescription to verification, ensuring that the parts developed will fit together in the virtual and hardware assembly.

A digital geometrical product description, a combination of all 3-D models of all relevant parts in their correct position in space is called a digital mock-up (DMU), and is used for the detection of collision between parts and for the simulation of the assembly process (Fig. 13.80).

To build up a DMU, it is necessary to have a bill of materials for the specific car. It can then be analyzed whether all part geometries necessary for a complete DMU have been provided by the engineering design or whether any part geometry are missing (Table 13.8). Since car development usually involves the development of a number of different car variants (left- and

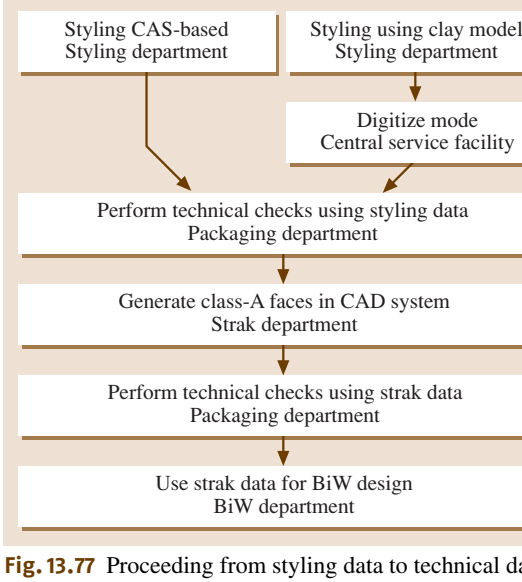


Fig. 13.77 Proceeding from styling data to technical data

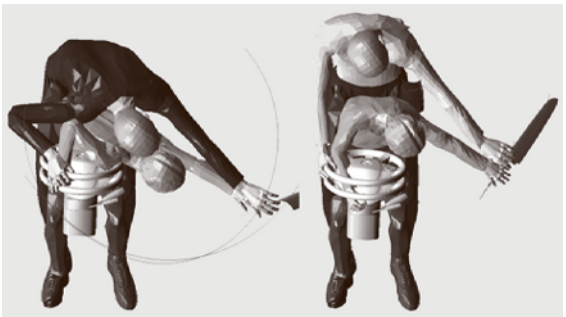


Fig. 13.78 RAMSIS application for verification of accessibility of door closing mechanism [13.82]

Fig. 13.79 RAMSIS application for verification of accessibility of rear door closing grip [13.82]

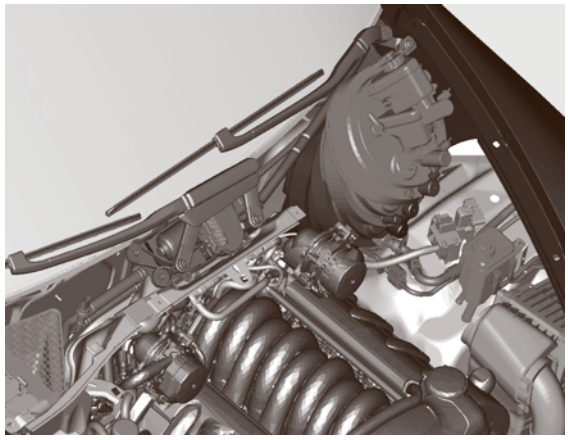


Fig. 13.80 DMU of components in the engine bay of a sports utility vehicle (courtesy of Dr. Ing. h.c. F. Porsche AG)

right-hand drive, automatic and manual transmission, different engines, and different customer-specific equipment), packaging has to define the most critical (in terms of space) car configurations, and will track the fitting of parts with respect to each other on the basis of these configurations. Based on experience, a trade-off is made between complete coverage of all car variants that are potentially critical and the expense of this level of coverage.

In order to be able to generate a DMU automatically, the designer has to assign his 3-D CAD model to a node in the bill of materials, and has to position it in each relevant car configuration. Positioning can be done either in absolute coordinates (relative to the vehicle null point) or in relative coordinates describing the position relative to another part on which it is, for example, mounted.

Part collisions can be checked on demand (when the designer wants to verify his part design in his specific geometrical surrounding), or on a regular basis as a complete geometrical car check. Conflicts between two parts are documented, and a recommended action is communicated by packaging to the two part owners.

Hardware. Most automotive companies use hardware to verify the space allocation of parts and the ergonomic suitability of the position of interior components. The engine bay, which contains flexible parts (cables, tubes) that cannot be appropriately modeled in CAD systems today, is one example. The seat box is another example, where hardware verification is used in order to check the ergonomic position of every element in the future seat. Both have their virtual counterparts, which are used earlier in the process.

Body Structure The main physical structure of a car is the body-in-white structure (BiW structure). With the layout of this structure, the basic performance features of a car are determined, too. To a certain extent, the BiW structure accounts for the weight of a car, the bending and torsion stiffness, and its crash behavior. If changes in the layout of the BiW structure are necessary later in the process due to conflicts in space management or unsatisfactory crash performance, these changes result in heavy additional cost, since changing this structure leads to necessary changes in all other sections of the car connected with the BiW structure and in the highly automated manufacturing lines in the factory which are planned, ordered, and built early in the process, too. Thus, methods for a thorough layout of this structure according to the goals defined for the car to be developed are used.

Virtual. Since body-in-white structures are developed in the early phases of car development, when no hardware is yet available, the focus of methods for layout of the BiW structure is on virtual methods. For the layout of a structure, generative algorithms can be used. One set of algorithms widely applied in the automotive industry is implemented in the package *SFE CONCEPT*. According to the definition of load cases, material coefficients, and major geometrical boundary conditions such as free space for the passenger compartment and trunk, the algorithm identifies at which points of space material is needed in order to fulfill the demands for smooth load flow and minimum deflection of the beams of the BiW structure. The result of this algorithm is a space structure of voxels which then can be used for a first layout of the structure in the CAD system. The result of this structure is then taken for finite element analysis to prove that the performance criteria are still met after the structure has been designed in such a way that it can also be manufactured.

Modal analysis of the BiW structure will show which frequencies will lead to resonance reactions, resulting in noise and discomfort for the passengers, and which therefore have to be avoided when guiding the load path into the BiW structure.

Depending on the car family the BiW structure is intended to cover, precautions for reinforcements in the structure have to be taken into account. An example is the need for reinforcement when the BiW structure of a coupé is intended for use in a convertible, since bend-

ing and torsion stiffness will decrease in this case due to the missing roof, in order to give the coupé additional stability.

Hardware. To ensure that the results of the calculation are correct, one of the first BiW structures to be manufactured as a prototype structure is often used for the verification of bending and torsion stiffness on testing machines. Since these tests are also quite inexpensive and the correct layout of the BiW structure is fundamental, this approach is reasonable. The first prototype with the new BiW structure, in most cases built shortly before the end of the concept phase, is used for first driving tests and, most important for the verification of the structure, in case of accidents. Due to the highly dynamic processes and its high dependency on dynamic material properties, the results of the simulation of a crash are still uncertain, which is why hardware verification is still needed before series development can start.

Aero/Thermodynamics The basic aero- and thermodynamic behavior of a car is determined by the layout of the exterior, the main air channels inside the car, and the position of the engine and other power-train components. Thus, this behavior has to be checked early in the process, when the styling process is not yet finished, in order to allow for necessary design changes to be made due to flaws in the aerodynamic performance [13.83–85].

Virtual For prediction of the aero- and thermodynamic behavior of a car, computational fluid dynamic (CFD) methods are used. With the help of these methods, the air flow around and inside a car can be visualized, a key performance indicator of a car – the c_w coefficient which contributes to the fuel consumption, especially at high speeds – can be determined, and air flow can be analyzed. Minimum air mass flows are defined, for example, for the cooling needs of the engine through the the radiator, and of the brakes.

The normal force to the road consists of the weight of the car and of the force generated by the air flow, the down force. The down force is also being calculated in the early phase, as it is critical for driving stability at high speeds. Changes in the exterior car shape and additional aerodynamic spoiler can affect the down force.

Hardware In order to analyze the airflow around and inside a car, flow analysis models can be used. Early in the process, scaled models can be used to evaluate exterior airflow and provide criteria to decide on a styling alternative that is more suitable in terms of aerodynamic performance. Later in the process, 1 : 1 scale models can be used for analyzing both exterior and interior airflow.

Fig. 13.81 Test with sports utility vehicle to verify performance in dry and dusty environment (courtesy of Dr. Ing. h.c. F. Porsche AG)

The hardware-based approach is losing favor against the use of virtual simulations.

Methods for Series Development

With the concept confirmed, the series phase starts. Increasingly virtual methods are substituting and being complemented by hardware methods.

A huge number of different development and testing methods are used in series development to ensure that

Fig. 13.82 Test with sports utility vehicle to verify performance in cold and snow environment (courtesy of Dr. Ing. h.c. F. Porsche AG)

Fig. 13.83 Test with sports utility vehicle to verify performance in wet environment (courtesy of Dr. Ing. h.c. F. Porsche AG)

Fig. 13.84 Test with sports utility vehicle to verify performance under extreme road conditions (courtesy of Dr. Ing. h.c. F. Porsche AG)

the components developed will meet functional, performance, and quality targets. Each automotive company has its own particular approach in terms of which methods are used for what kind of evaluation at which point in time. Figures 13.81–13.84 give an overview of different tests to be performed in the development of a sports utility vehicle.

Using three example methods, a general insight into the methods for series development is provided below.

Climate Simulation. A climate chamber (Fig. 13.85) is used to simulate extreme climate conditions. Heat and cold are simulated, together with different levels of humidity and sunlight incidence. Different kinds of behavior are analyzed in the climate chamber: car functions such as door opening, and heating and cooling of the passenger cell are analyzed, as well as engine start-up performance under extreme conditions. An extension of the climate chamber is the climate wind tunnel, in which different wind flow can additionally be generated with the car standing still or using its engine to turn the wheels on rollers mounted to the floor.

Climate simulation on the test track is usually combined with analysis of the behavior of the car and especially its cooling system at high speeds. To simulate the car being parked in a garage with a hot engine, after these tests the car can be put into a closed chamber to simulate this extreme operating condition.

Climate simulation is also done in extensive tests in different climate zones in the world such as Northern Canada and Death Valley. Here, it is important to minimize the effort in transporting the test cars between the different test locations while maximizing the difference in climatic conditions, in order to minimize cost and time lost during transportation. Snow, sand, different road conditions, and stop-and-go traffic in metropolitan areas are also covered during these tests.

Chassis Tuning. For tuning of the chassis system different driving situations are simulated by test drivers. Additional adaptations in terms of springs, dampers and, most important, electronic stability systems are derived. In order to cover all combinations in which a car can be driven by a customer, the variants of tire size and type, usage of winter chains, and the different engine, chassis, and body types have to be taken into consideration. Therefore, with the large number of car variants that a customer can order, the effort required to tune the chassis can increase greatly. Table 13.9 shows a sample combination matrix for chassis tuning tests.

Fig. 13.85 Car test in climate chamber (courtesy of Mercedes Car Group)

Table 13.9 Sample combination matrix for chassis tuning tests

Tire size	18″	19″	20″	
Tire type	Summer	Winter	All season	Off-road
Winter chain	Without	With		
Engine type	3400 ccm	4500 ccm	4500 ccm turbo	
Chassis type	Normal	Comfort	Sport	Adaptive
Car body type	Small	Wide		

Besides tuning of the chassis, it also has to be verified that there is enough space between the tire (all potentially usable tires) and the fender liner under all driving conditions. For this purpose, early in the process wheel covers are generated as boundary faces for the movement of the tire under every condition in the CAD system, on the basis of which the fender liner is designed. The correct sizing of the fender liner is then verified in hardware tests later on. Most critical in terms of space needed are large tire sizes with a sporty chassis layout and the usage of winter chains.

Hardware-in-the-Loop. Hardware-in-the-loop (HIL) is an approach to test certain hardware components in their future operating environment, when this environment is not already available, i. e., when other interoperating components are not available yet. In this case, the other components as well as the environmental conditions are simulated by a computer program in real time, and the simulation parameters are input to the component to be tested. HIL is used for mechanical, electronic, and electromechanical components.

The HIL approach is used to decouple the behavior of components such as controllers in an interconnected network (for example, the CAN bus). With clearly defined interfaces between the components and the definition of the communication bus structure, the HIL method is similar to the methods used in software engineering for some time already.

An example of the utilization of HIL is the testing of controllers for active body control in a test-bed with simulated signals from sensors and actuators before the controller is put into the networked car environment in a prototype.

Tests of a component can be reproduced, and thus also critical conditions in terms of operating behavior can be reproduced and used as test criteria for approving the release of the component.

With the decoupling approach of HIL, product development time can be reduced, since it is not necessary to wait for the last component in a network to be available in hardware. Also, due to the modular structure of the network of components, the complex interaction of components can be controlled.

Cross-Functional Methods

Product Structure and Bill of Materials. The car is logically and/or functionally structured. The power train, the doors, and the cockpit might be high-level elements of this structure which are substructured further. The leaves of the structure tree represent the parts or components. The part or components describing data is matched via PDM software and transformations that describe the position of the specific part or component are added. The functionality of the PDM software enables the digital generation of assemblies or the entire car. For this purpose the product structure will be interpreted, and the geometrical description of the parts and components can be loaded and positioned element by element using the attached transformations. This functionality is the basis for all kind of simulations (Fig. 13.86).

The product structure again is the basis for the BOM (Fig. 13.87). In the bill of materials the leaves of the structure of the car are focused upon. The geometrical description itself is neglected; its existence and the level of quality are of importance and, hence, represented. Instead, further information is added which enables the configuration of the car. The BOM describes the contents of the accessory packages to be offered, the specifics of the American, European or Japanese versions, the colors that will be available, the options that can be chosen, and the dependencies that exist between these configurations.

The BOM, and the system in which it is handled, form the basis for all logistical activities. Material and supplier parts are ordered using the BOM for prototypes as well as the final, customer-specific car. The plants in which the cars will be assembled can be planned with respect to capacities, required tools, robots etc. and manufacturing costs can be calculated since the part-

specific costs are also added in the BOM, to name just a few applications of the BOM.

Since the role of the BOM is essential, its quality is a major aim in car development.

Data Quality. Since the virtual models of parts, components, and assemblies are used for diverse simulations (see *Methods for Product Layout and Conceptual Development*) the representation of the real pendant has to be of high quality. Only by guaranteeing high data quality can costs be effectively saved by substituting physical prototypes by virtual ones early in the process.

The parts have to be *technically* mature, which means they have to meet the requirements by fulfilling the function specified and have to be manufacturable. With respect to the car, they have to fit in the system and their assembly has to be possible. On the other hand the parts have to have a high quality with respect to *formal* demands. For example, it is important that guidelines, both general and company specific, have been taken into consideration in order to prevent unnecessary correction iterations. The parts have to be described completely and all required extended information has to be provided. This information completes the geometrical description of the part, component or assembly and gains importance during development process.

Examples are material, weight, center of gravity, tolerances as well as metadata such as the person in charge of the part, the supplier name or in which context the part will be used or assembled (depending on the type of car, accessory package, or options to be chosen by the customer).

The responsibility for the quality of data belongs to the part owner. To achieve technical quality the part owner is supported by departments that investigate the packaging (see *Packaging and Ergonomics*) and digitally secure the assemblability. Companies often use tools to check the quality of the geometrical data, particularly when data comes from suppliers. These tools check whether planes are closed, parameters are set, and so forth. The formal quality can be controlled by monitoring tools which check specifically defined parameters. Sometimes the part owner is actively supported in completing the nongeometrical information by departments that are in charge of product data management. These departments finally check the data set of a part, component, and assembly. The positive result of a check leads to the release of the part, component or assembly. This release also means that the digital model is reliable and can be used in further development (see the following section).

Release. The importance of high quality of data has already been discussed (see *Data Quality*). The measure by which the required quality of data will be

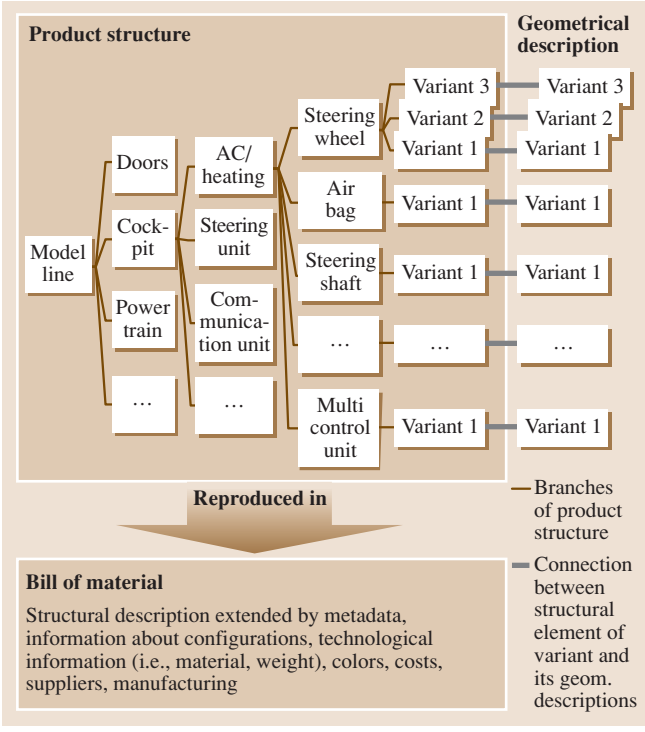


Fig. 13.86 Connection between product structure and virtual model

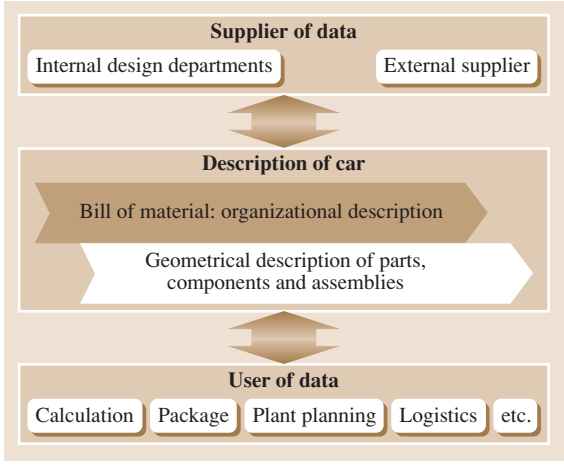


Fig. 13.87 Central role of the BOM in the car development process

Table 13.10 List of requirements to be checked before starting a release process for a part/assembly (not all aspects are always appropriate)

Weight targets fulfilled	Cost targets fulfilled
DFM checked	FMEA carried out
Relevant laws and regulations checked and fulfilled	Standards checked
Carryover part contained	Production process capability proofed
Safety aspects taken into account	Part assigned to car variants
Open points from design reviews settled	Tests absolved positively

Table 13.11 List of general aspects to be considered before starting the release process

Packaging	Styling
Calculation and simulation	Materials laboratory
Acoustics	Thermodynamics
Aerodynamics	Safety
Process planning	Patents
Purchasing	Supplier
Testing	Homologation
Production	Quality
Sales	After sales
Controlling	

finally checked is the release, which concludes the development-specific activities at a certain version of the part, component or assembly. The label "released" for an object makes the virtual model visible for downstream processes such as design of tools, logistical processes, and planning of manufacturing units and plants. The released model is guaranteed to have the required level of maturity and quality in order to start hardware processes. Hence, every part which is needed to be manufactured and/or assembled has to be released. Only released parts can be ordered, regardless of whether the part is used for the production of a prototype or the final assembly of the car.

The release processes differ slightly from OEM to OEM. Nevertheless a general process can be recognized. The release process described below is a generalized example which demonstrates the major characteristics of the process.

The prerequisite for the common release process is the completed design of the part, component or assembly. It has to fulfill its function(s) and meet all defined requirements at the specific point of development. The geometrical description has to be complemented by all required information. What this information is depends on the point of development; for example, only a few pieces of technological information, such as tolerances and connecting details, are required in the early phases, whereas the weight and the center of gravity have to be available as precisely as possible from the start of design.

The release itself requires a positive evaluation during several steps. An overview of the process is presented in Fig. 13.88.

Design is *inventing* the part and secures its function. Before the release process is started several aspects should be taken into consideration. Tables 13.10 and 13.11 show examples of these aspects.

When the part has the required maturity and shall be released, the person in charge of the part has to lock the model in order to prevent further manipulation. At this time the required data quality has to be secured by design or related departments. Data has to be checked by sight or specific tools for the aspects to be considered (Table 13.11). Depending on the process specifications and the specific release level during the car development process these aspects differ with respect to the expected maturity of data (Fig. 13.82).

The locked and prechecked part will then be passed onto the *drawing check*. Now the model can, for example, be checked for:

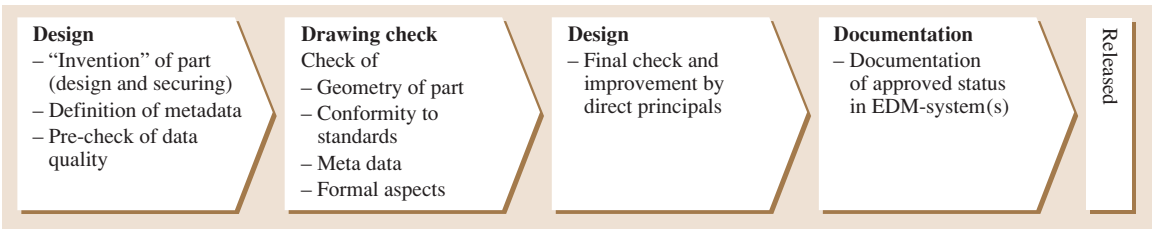


Fig. 13.88 General proceeding of a release

- The manufacturing specific design of the part
- The conformity of geometry, material etc. to standards
- The existence of the required technological data
- The existence of the required metadata
- The existence of the required attributes

After a positive evaluation of the draw check, the part and its evaluation results are *transferred* back to the design where the persons responsible for design have to confirm the evaluation from their point of view. Aspects that have to be considered by the person(s) representing the specific design department where the part was created include whether the part is technically acceptable and whether recent decisions prevent a release of the specific part.

When the confirmation release is successful this information is documented in the engineering data management (EDM) system. The part, component or assembly is now labeled as "released". This attribute is also transferred to the BOM. In this way downstream processes can now access the geometrical description of the part, and the release process is completed.

The various forms of releases are differentiated in Fig. 13.89.

Car development is a complex process in which suppliers develop a large share of the car. The parts, components or assemblies will be designed at the OEM or be provided to the OEM by the supplier. Although the supplier naturally checks the required maturity for the development to be delivered, the release of the data is done by the OEM in order to achieve uniform data quality supplied to the process.

Release without application	
→ New part	(general design, i.e., standard parts or generally applicable parts like tubes, clamps, surgical gloves, washer fluid)
Release with application	
→ New part	(specific design, i.e., body parts or the specifically adjusted tube)
→ New application of existing part	
→ Change of design	(part has to be replaceable in all applications)
→ Change of level of maturity	(no changes of design have been made, development phase changes by definition)
Examples	
⤷ Part looses validity	(for a specific application, another application might stay valid)
⤷ Exchange of part	(for a specific application)
⤷ Change of relevant equipment package	
⤷ Change of quality	

Fig. 13.89 Kinds of release

Not covered by this description of release processes and particularly challenging is the release of software. Since all OEMs handle the securing of software maturity differently and the processes are often handled very flexible it is not currently possible to describe a representative process.

As mentioned above the aspects to be checked in a release depend on the specific milestones of the car development process of the particular OEM. Figure 13.90 shows some examples of the form of a release with respect to maturity of car development.

Prototype and Test Management. The different departments in car development have special needs for tests. During tests, certain aspects of the behavior of the car are analyzed. In order to achieve the most efficient utilization of prototypes, tests are scheduled and assigned to prototypes in a way that minimizes the number of prototypes and the effort required to rebuild prototypes, while maximizing the usability and special needs of prototype equipment and maturity for the design departments. Prototype and test management is usually a central facility that covers all car development projects (Fig. 13.91).

Weight Management. The weight of a car is a basic layout parameter and influences fuel consumption, and the distribution of the weight between the front and rear axles of the car is a parameter that influences the driving performance of the car.

Therefore, the definition of weight targets and weight management during the car development process is important to ensure that the car will meet the expected performance criteria.

In the target building process, the weight target for a car is defined by:

- Benchmarking of car weight of competitors in the market segment
- Key performance parameters of the car to be met (such as fuel consumption)
- Top-down definition of targets by proportional ratio of weight from the predecessor car
- Consideration of negative and positive weight contribution of new components (such as infotain-

ment components) or technologies (aluminum body structure)

Weight targets are broken down from the whole car into modules and components to parts, and the designer has to meet the weight target while keeping the cost and time targets as well as the function in mind.

As soon as weight targets are defined and agreed upon, weight management is started to control the current status of the car weight at each step of the car development process. The quality of the weight status can be divided into three different categories. In the early phases, when there is no 3-D part geometry available, the weight declaration is in the state of *estimate*. With engineering judgment, the responsible designer will feed this value into the weight management process.

As soon as a reasonable 3-D part geometry is available and the material is defined, the weight can be *calculated* by using CAD functionality for calculating the volume times the material density. The third level of weight status is reached as soon as parts are available in hardware. The weight can then be measured using a scale. The quality of the weight state *weighed* depends on whether the parts are made out of series material and series tools or whether prototype material and tools are used, and eventually corrective factors are used to forecast the final part weight (Fig. 13.92).

During most of the car development process, there will be parts concurrently in the state of *estimated* and *calculated* or *calculated* and *weighed*, making the overall weight state a combination of both states until near the start of production (SOP).

Weight is usually tracked in the BOM.

Weight targets are defined for a number of different car variants, and these have to be checked continuously. Finally, it is important to meet the needs of homologation of a car in the different countries to which the car will be exported, and the different regulations for measuring the car weight (tank full or empty, supposed number and weight of passengers).

Change Management. The dynamics of car development does not end with its introduction to the market. A car will continuously be developed further. On average each part of a car will be changed approximately three times, and engine parts about eight times. Up to the point that a car model line ends, the number of spare parts increases by a factor of five.

Changes can become necessary because of field feedback (guarantee and warranty, quality, customer ex-

Proceeding of product development

Kind of release	Release for planning	Release for tools	Release for purchasing	Release for disposition
Prerequisites	Part design is mature for testing	Results of testing positive for step-by-step release	Part desing finished, test results are sufficient	Evaluation by production is positive
Activities before starting release process	– CAD model created – Part information defined – BOM adapted – Signed by principals	– CAD model, BOM and part information changed due to results from testing – Signed by principals	– Part design finished – BOM and part information finally adapted – Signed by principals	– Sample delivery and testing positive
		Release		
Activities triggered by release	– Evaluation in terms of DFM by production – Definition of production methods and tools – Purchasing of prototype parts – Disposition for tests – Check series delivery capabilities of supplier	– Generation of manufacturing BOM – Order of tools for series production – Finish planning for logistics processes	– Adaptation of manufacturing BOM – Planning of SOP – Planning of production capacity – Securing of ramp up of series production by contrasts – Purchasing of pre-series parts	– Contract for unlimited part delivery according to volume planned

Fig. 13.90 Kinds of release with respect to maturity

pectations, requirements from sales), measures due to legislation, optimization (costs, production, and quality), and technical development.

Change requests always have their reason. Nevertheless not each request can be realized since each change requires enormous effort, as a complex process is necessary to realize a change. An idea of what has to be carried out to implement a change is in a new car is presented in Fig. 13.93.

When a change request (Fig. 13.93 (4)) is generated for a model line that is in production (3) the change process starts in development again (5). Although the process of adaptation (6) differs slightly from the general development process, the main steps are the same as shown in Fig. 13.93 (6). When the necessary adaptations of production tools and proceedings are implemented the changed car (8) can finally be manufactured (7).

The entire process of realization of a single intention to change costs about 20 000–50 000. This is the reason why a change request has to be considered carefully. It has to be considered whether the reason to change specific features of a car is profitable enough to cover the costs resulting from the process necessary for its realization.

In Fig. 13.94 the main steps to evaluate an intended change are shown. The two-level release process helps to extract unnecessary or unprofitable change requests early in the process. When the change request or intention is described by an expert, for example, a principal of the specific design department can evaluate it. This evaluation functions as an early filter, when only little effort is invested. An approval leads to a phase of detailing of the design of the intended change. The de-

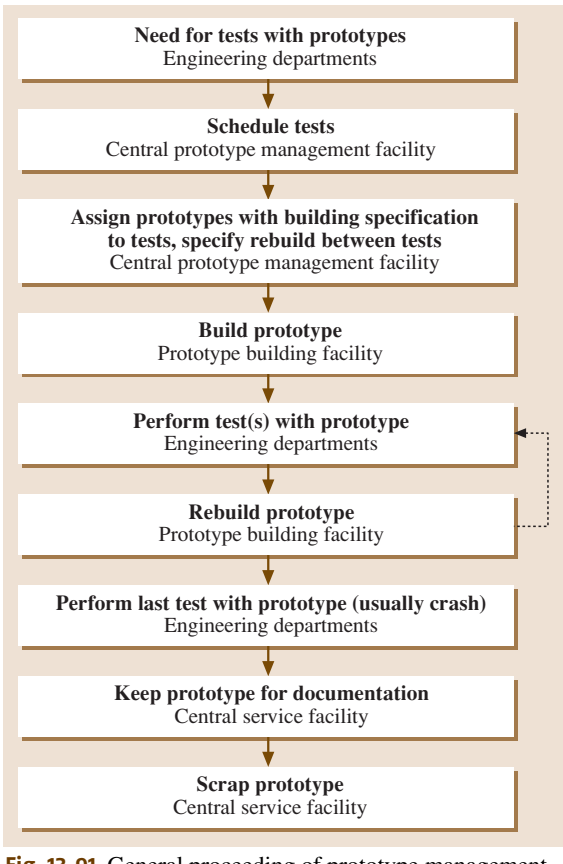


Fig. 13.91 General proceeding of prototype management

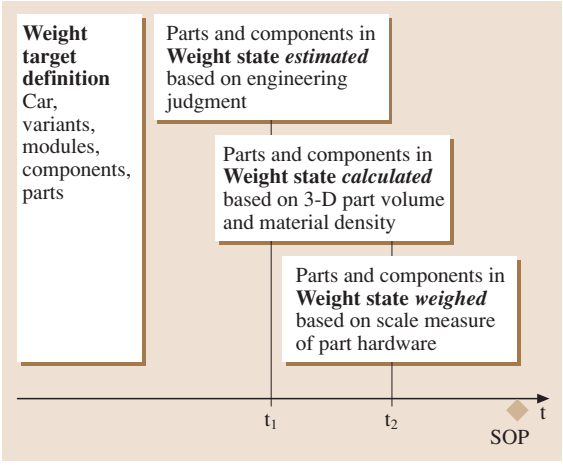


Fig. 13.92 Total weight state based on different part weight states (examples t_1 and t_2)

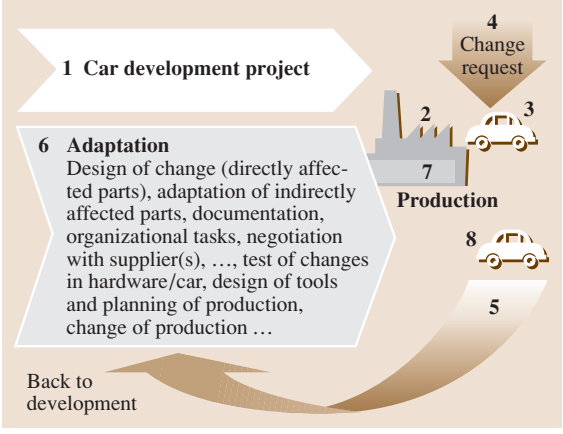


Fig. 13.93 Effort resulting from the realization of a change intention

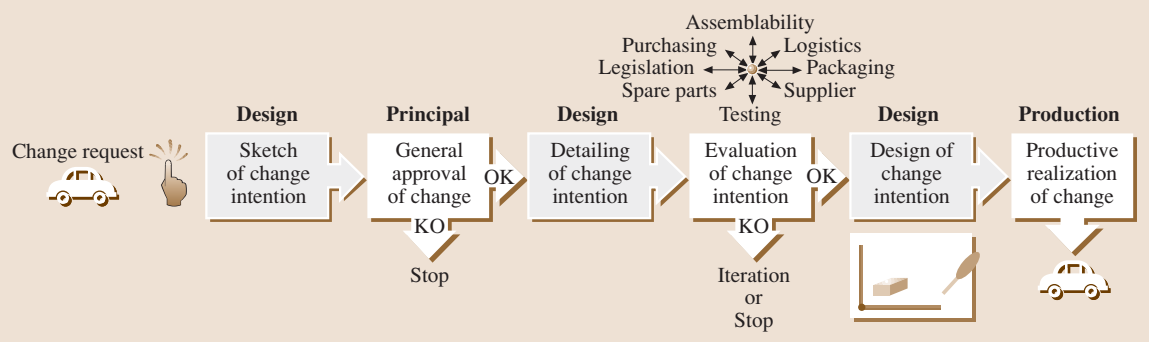


Fig. 13.94 Example of a change approval and realization process (after CAD/CAE in Automotive Industry, Technical University of Berlin, Institute of Vehicle Technology) (KO = knock-out)

tailed change request is then evaluated by all affected departments, as shown in the figure. Due to the more detailed specification an evaluation of the intended change with respect to affected topics by the experts is now possible. When these experts approve the request the realization of the intended change starts with the design.

The process of the evaluation and realization of a single change itself causes an enormous effort. However, during the car's lifecycle and in the various departments involved in car development many change requests are generated. Since there are about 8000 changes in one car model line and approximately 50% from that after start of production, one can imagine that it is not possible to realize each change intent in the production line once it is completely defined. A change in production might, for example, request a standstill of the line. This is expensive and must be coordinated so that many changes are carried out at the same time.

Another aspect which needs coordination is the testing of changes in hardware. In order to keep costs low several changes should be tested in one production vehicle used for testing series aspects. A prerequisite is that the changes do not affect one another. Hence, the realization of changes has to be planned in detail.

Integration of Development Partners. Tier 1 suppliers today have in-depth knowledge of their technologies, components, and their application in a car which auto-

Table 13.12 Nonengineering functions to support car development

Purchasing	In the early phase of the car development project, engineering partners and system suppliers who collaborate in concept development are sourced by purchasing. Contracts are made about the content of engineering service provided up to the provision of prototype parts for first hardware tests. Afterwards, most automotive companies have a global sourcing process in order to identify which supplier will provide the parts for series production and select this supplier on the basis of reliability of supply and price.
Marketing	Marketing represents the voice of the customer. It is mainly involved in the definition of the car development project (positioning of the car to be developed in the market, definition of target price, definition of options to be offered).
Controlling	Each car development project is measured by a number of different parameters, one of the most important of them cost. In most automotive companies, budgets are assigned for each simultaneous engineering team responsible for a specific area of the car. These budgets consist basically of engineering cost (internal and external), prototype material cost, and investment for sourced parts and for production machinery. Controlling keeps track of the predicted budget consumption against real consumption, and provides valuable information to the team for taking necessary corrective actions to meet the cost targets.

motive companies usually can no longer afford to have due to the extreme specialization required for the different technologies. Therefore, the integration of these suppliers of the car development project from the very start is of vital importance. Examples of technologies and components developed by suppliers are headlamps and electromechanical components such as antilock braking systems.

When integrating suppliers into car development, one has to consider:

- The product substance
- Its extent in terms of engineering service
- The interaction between the supplier and the OEM

Depending on this, the following aspects have to be agreed upon in contracts with the supplier:

- The kind of data exchange (mail, direct access to systems from the outside or by resident engineers)
- The kind, form, and contents of the results to be delivered by the supplier
- The kinds of simulations and tests to prove the suitability of the concept and its application in the context of the car
- Delivery of prototype parts, if applicable and necessary
- Delivery of parts for series production, if the supplier will also source the component for series supply

In order to integrate development partners successfully there has to be a trade-off between open transfer of information and retaining specific expertise in the OEM and the relevant supplier.

Nonengineering Support of Car Development. During car development, the engineering departments are supported by various nonengineering functions in the company. Table 13.12 briefly describes the most important functions.

13.3 Railway Systems – Railway Engineering

13.3.1 General Interactions of Modules of a Railway System with Surrounding Conditions

Railways have many technical and economical interfaces to the surrounding world, as indicated in Fig. 13.95. The aims of the railway are usually provided externally, from policy and economics regarding market, finances, and environment. These aims are transformed into strategies be the management of a railway company, defined by instructions to several subareas such as marketing, which define the product in terms of timetable and comfort. The timetable provides lots of information; it defines the locations to be connected and the distances to be overcome. By defining the times of departure and arrival, the travel speed is fixed. Also the frequency of operation of trains is defined.

The railway operation must be able to fulfill this requirements by providing adequate, educated staff in trains and at fixed locations. Energy to move the trains must be provided at the right locations. Communication must be enabled over one or even several lines with many trains. If the schedule fails, disturbance management must be able to restore the system to proper operation as soon as possible, whether failure is caused by exterior or interior reasons.

Infrastructure such as the tracks, perhaps catenary, the design speed of the track and its gage, the structural gage, the axle loads, stations for passengers and/or loading/unloading facilities for goods must all fit the demands. Information technologies for passengers are also gaining in importance.

The type of vehicles chosen must be adequate for the required operation, for instance, locomotives, coaches or diesel multiple units (DMUs). The vehicles must fit the infrastructure and the operation in terms of speed, axle load, etc..

Maintenance must provide reliable system elements by avoiding failures because of the effects of wear. Maintenance is increasingly being outsourced today.

If all of the elements shown in the boxes on the diagonal in Fig. 13.95 are provided by one company it is called an integrated railway otherwise it is known as a segmented railway.

The interaction of the elements produces results in terms of earnings, and the quality of the process (for example, punctuality).

Because of high cost pressures the aim for a economically sound railway system is to run as fast

[(km/(vehicle of the fleet × day)] and as reliably as possible.

A high number of kilometers per vehicle in the fleet is important, as the costs for vehicles, stationary equipment, and operational staff are a function of time, whereas the earnings from passenger and freight traffic is more or less a function of distance. Also fast running in form of fast point-to-point transportation is itself attractive for passengers and goods.

Good reliability reduces the number of spare vehicles and spare staff and also reduces operational risk. Of course this is an important quality measure for clients in passenger and freight operation.

Rail vehicles must therefore be designed according to these requirements.

Duration of Passenger Exchange

As stopping time has a negative effect on the aims mentioned above (covering as long distances as possible per vehicle and day) stopping time should be as short as possible, which results in many technical and operational constraints.

As an example the different elements or modules define the stopping time of a train (Fig. 13.95).

The duration of stops (intermediate stops and in the terminus/stations) is affected by the following parameters:

1. The height difference between the platform and train level, which should be as small as possible.
2. The gap between the train entrance and platform, which should be as small as possible (the ideal being gapless high-level platforms in a straight line and without curves).
3. The ability of passengers to handle the process; large amounts of luggage and disabled persons slow down the loading and unloading process. Disabled persons are for instance small children and older persons.
4. Duration of the door opening and closing process. Modern semiconductor-controlled doors often have very lengthy closing procedures, for instance, 20 s for the InterCity Express 1 and 2 (ICE 1/2) trains in Germany, and similar values for the *train à grande vitesse* (TGV) in France. The reasons for this are the slow processors used for door control and the connection to the vehicle bus.
5. Number of doors. The number of doors should not be to small because of the possibility of door failures, but for long-distance traffic doors are costly because of the door costs itself and because the space behind the door is not usable for passenger seating space.

Especially for trains with frequent stops it is very efficient to have short stopping times. For sure the preprocess and postprocess time are not productive at all and the passenger exchange time is the more effective the shorter the time is for a given number of exchanging passengers (Fig. 13.96).

Lifecycle Costs

The annual distance traveled by rail vehicles is rather high, between 100 000 km/year for tram cars, over

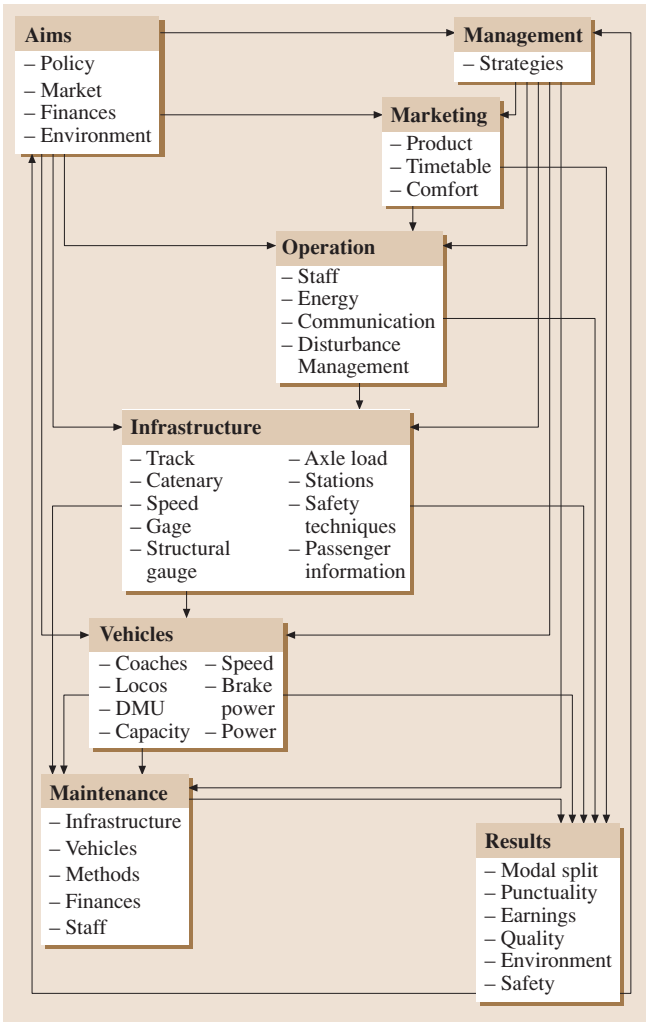


Fig. 13.95 Railway transportation – a system with strong interference

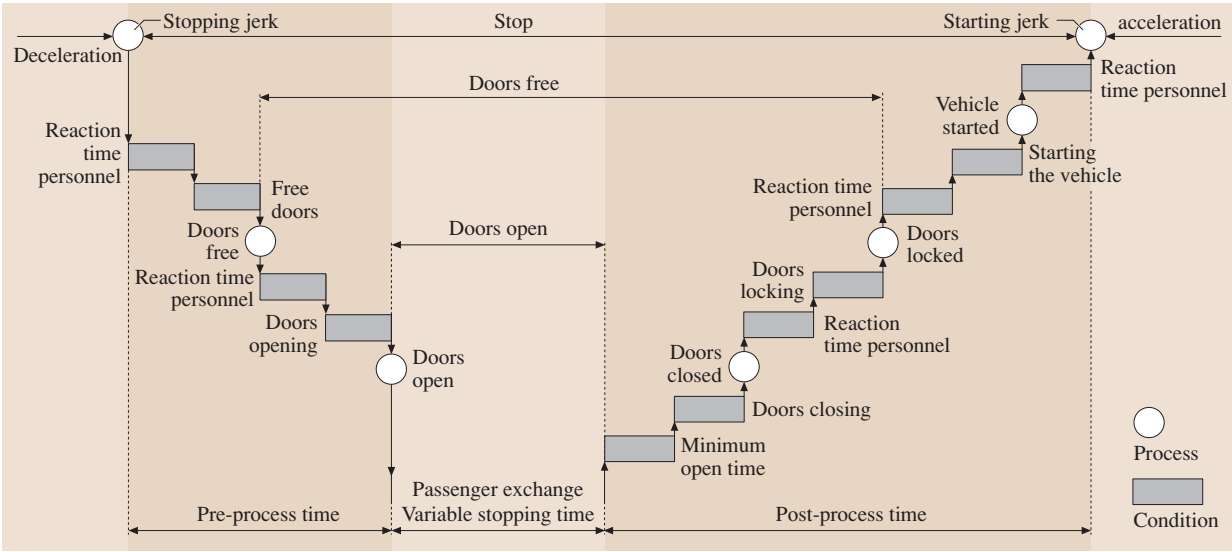


Fig. 13.96 Stopping, door opening/closing, and passenger exchange (after [13.86])

200 000 km/year for regional trains and freight locomotives, and up to 500 000 km/year for high-speed rolling stock. For freight cars it may range between 10 000 and 200 000 km/year. Lifetime is typically 30 and sometimes 40 years. Investment costs are rather high and typically cover one-quarter to one-third of the lifecycle costs. Other costs are maintenance and operation. In countries where wear- and noise-related track prices are charged, at least the difference between the cheapest vehicle and the actual vehicle must be respected in the lifecycle cost as a new component.

Reliability, Availability, and Safety

The basis for reliability, availability, maintainability, and safety (RAMS) is the European standard EN 50126 *Railway applications – The Specification and Demonstration of Reliability, Availability, Maintainability and Safety* (September 1999) (Fig. 13.97).

RAMS is needed due to the increasing complexity of railway systems and to achieve a high-quality process. In a competitive environment (mainly from other modes of transport such as road, air, and even water) this is essential to survive economically. A very reliable transport (for instance, 99% or even more trains on time) with high speeds and therefore little standby time must be arranged. Good reliability reduces the need for spare units and spare personnel; high speeds increase productivity as payments are received on a basis of passenger-km or ton-km measures, whereas the expenses for equipment, both trains and infrastructure, and also personal occur on a time basis.

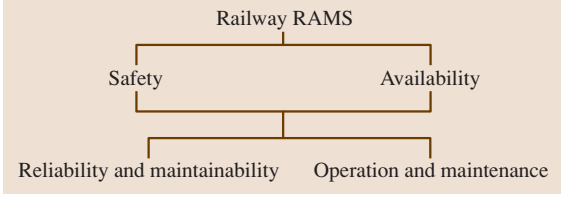


Fig. 13.97 Interrelation of railway RAMS elements

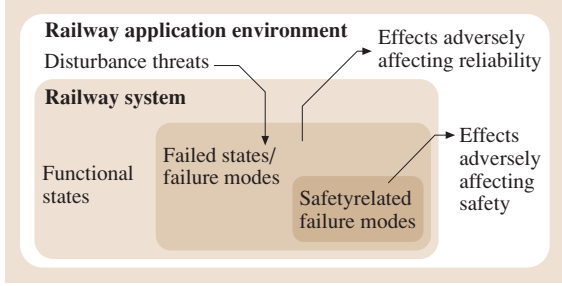


Fig. 13.98 Effects of failures within a system

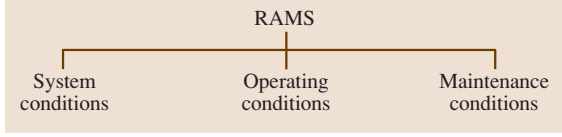


Fig. 13.99 Influences on RAMS

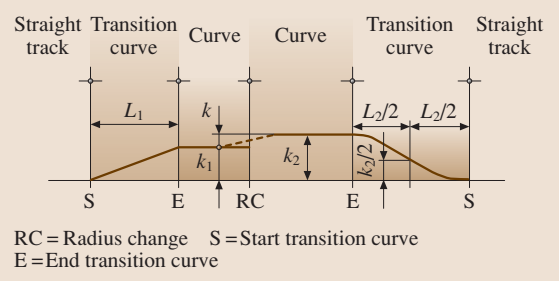


Fig. 13.100 Transition curves and ramps shown as curvature $k = 1/R$

Here mainly the process is discussed; targets are only mentioned as an example. RAMS covers the whole lifecycle of a system, and must be followed by the railway authorities (track and train operators) and the railway industry (system houses and suppliers).

The following definitions are used (Fig. 13.98):

- *Reliability*: the probability that an item can perform a required function under given conditions for a given time interval (t_1, t_2).
- *Availability*: the ability of a product to be in a state to perform a required function under given conditions at a given instant of time or over a given time interval, assuming that the required external resources are provided.
- *Maintainability*: the probability that a given active maintenance action, for an item under given conditions of use, can be carried out within a stated time interval when the maintenance is performed under stated conditions and using stated procedures and resources.

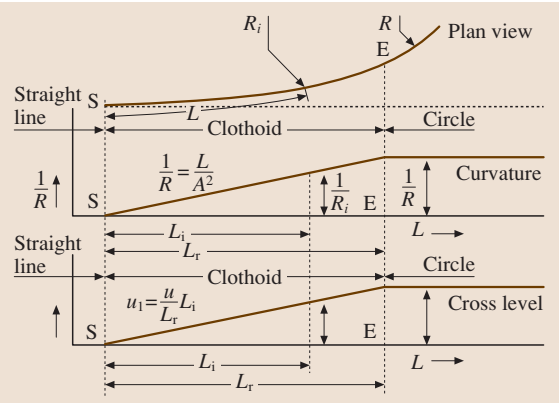


Fig. 13.101 Transition curve with linear ramp, *top*: plan view, *middle*: curvature, *below*: cross level

- *Safety*: freedom from unacceptable risk of harm (note that safety does not mean that no accident occurs at all).

Risk is defined as the product of size of the resulting destruction and the likelihood of the occurrence of the event. This likelihood of occurrence may be reduced by huge safety margins or by diagnosis that improves the probability of the recognition of a defect at an early and still not dangerous stage.

Around 1980 railways demanded an availability of about 80% for their fleet. Some railways operating under difficult conditions, for instance when suffering from the requisition of spare parts for cannibalization for other units, often did not reach 50%. Today figures of 95% for electric rolling stock and 92% for diesel rolling stock or even higher are demanded.

The definition of availability does not include accidents (for instance, at points during shunting) and vandalism (for instance, broken windows or broken chairs due to rioting). Operational retardation is also not included. Because of all of these reasons the number of spare units which are neither available nor down is greater than zero.

The basic equation for availability is

$$\text{Availability} = \text{MTBF}/(\text{MTBF} + \text{MDT}) \,, \qquad (13.2)$$

where MTBF is the mean time between failure, and MDT is the mean downtime.

Technical concepts of availability are based on a knowledge of:

1. Reliability in terms of:
   - All possible system failure modes in the specified application and environment
   - The probability of occurrence of each failure or, alternatively, the rate of occurrence of each failure
   - The effect of the failure on the functionality of the system
2. Maintainability in terms of:
   - Time for the performance of planned maintenance
   - Time for detection, identification, and location of faults
   - Time for the restoration of the failed system (unplanned maintenance)
3. Operation and maintenance in terms of:
   - All possible operation modes and required maintenance, over the system lifecycle
   - Human factor issues

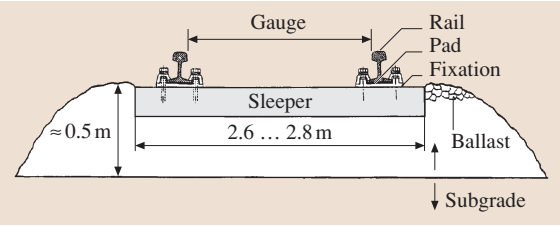


Fig. 13.102 Ballasted track bed (Table 13.13)

Table 13.13 Explanation for Fig. 13.102

Element	Function (construction)
Rail	Guidance of the vehicle, load distribution (rail 1 : 20 or 1 : 40 tilted)
Sleeper	Load distribution
Ballast	Load distribution, elastic element and damping device, distribution of the rail forces in all three directions, noise absorption of rolling noise and gearbox noise

RAMS is influenced by the system conditions (vehicle and track, mainly influenced be the vehicle and track producers), operating conditions (the knowledge of the onboard personal and the available data), and the maintenance conditions (knowledge and equipment of the maintenance personnel and the maintenance facilities) (Fig. 13.99).

Methods to reduce downtime include:

1. Improving information to reduce inspection time by diagnosis
2. Parallel maintenance processes, including simultaneous processes of repair, inspection, refilling of water, sand, fuel etc.

For instance, cleaning of the interior (and probably the exterior) and toilets should be done simultaneous with refilling of water, sand, and probably fuel (while ensuring safety), a small overhaul, and module exchange (wheelsets, drive components, air-conditioning modules, etc.).

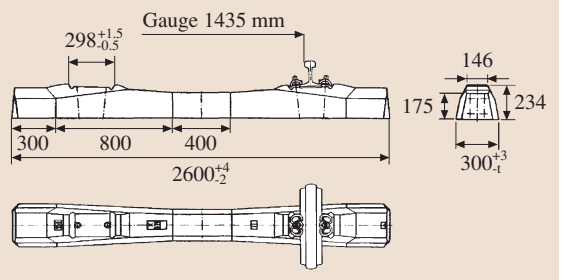


Fig. 13.104 B70 concrete sleeper, weight 445 kg (Pfleiderer Infrastrukturtechnik GmbH & Co KG, Neumarkt)

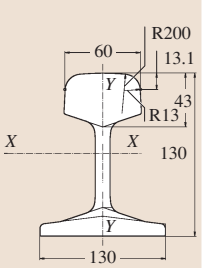
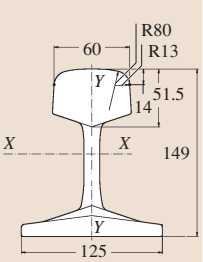
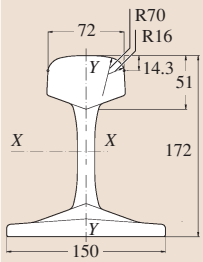
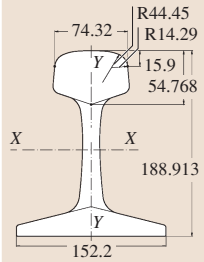
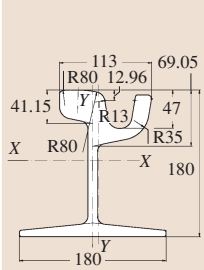
Rail profile	VST36	49E1 (S49)	60E2	AREA141AB	60Ri2 (Ri60-13, Ri60N)
					
	Flat bottom rail	Flat bottom rail	Flat bottom rail	Flat bottom rail	Grooved rail
Mass (kg/m)	35.82	49.39	60.05	69.88	59.75
Area (cm^2)	45.63	62.92	76.5	89.02	76.11
Moment of inertia $X-X$ (cm^4)	1009.3	1816.0	3022	4180.0	3298.1
$Y-Y$ (cm^4)	157.7	319.1	511.3	620.7	920.1

Fig. 13.103 Common rails with specific data. Rail profile, *TSTG-Profile-Handbook* (TSTG Schienen Technik, Duisburg)

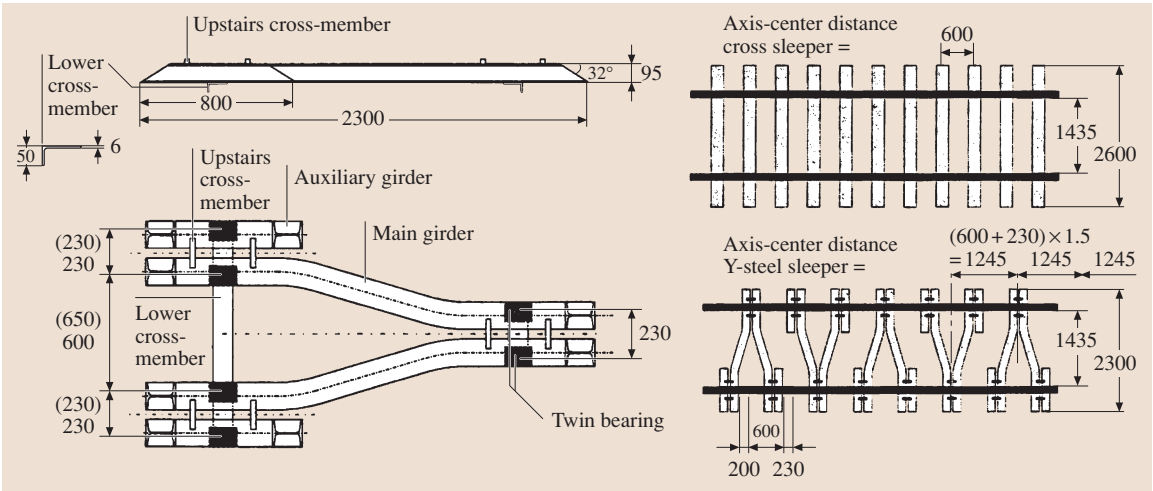


Fig. 13.105 Y-steel sleeper and comparison of steel, wood, and concrete sleeper track beds [13.87]

Simultaneous maintenance of fix train-sets (the locomotive plus wagons) and not locomotives separately from wagons avoids uncoupling and coupling time.

Systems diagnosis is becoming a major issue, and technical systems are becoming increasingly complex. Fault conditions are more difficult to reproduce, but fault analysis times should be reduced. Diagnosis systems for rail vehicles should have three levels:

1. Diagnosis for train personnel, to provide information for greater availability, and advice regarding redundant system operation
2. Diagnosis for maintenance personnel, to provide direct advice for required maintenance operations
3. Diagnosis for technical management, to provide data for reliability statistics as a basis for system improvements or at least spare parts management (to enable the greatest reduction of spare-parts storage)

Two requirements of the diagnosis system that must be fulfilled are:

1. All data must be collected in one system for the whole vehicle (not separate systems, for instance, for the diesel motor, drive, doors, toilets, etc.).

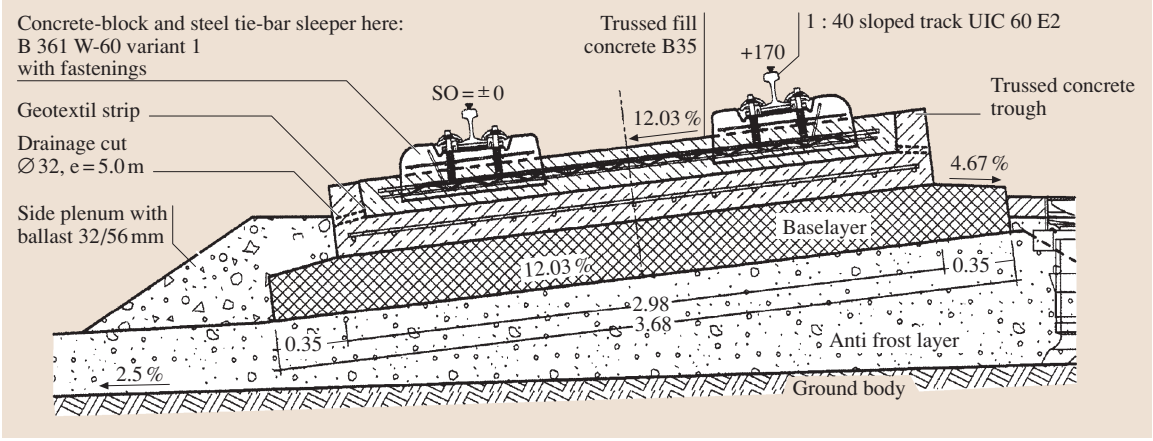


Fig. 13.106 Ballast-less track at the Cologne–Frankfurt high-speed line in a curve with maximum superelevation of 170 mm (Walter-Heilit, Munich) [13.88] (UIC = union internationale des chemins de fer)

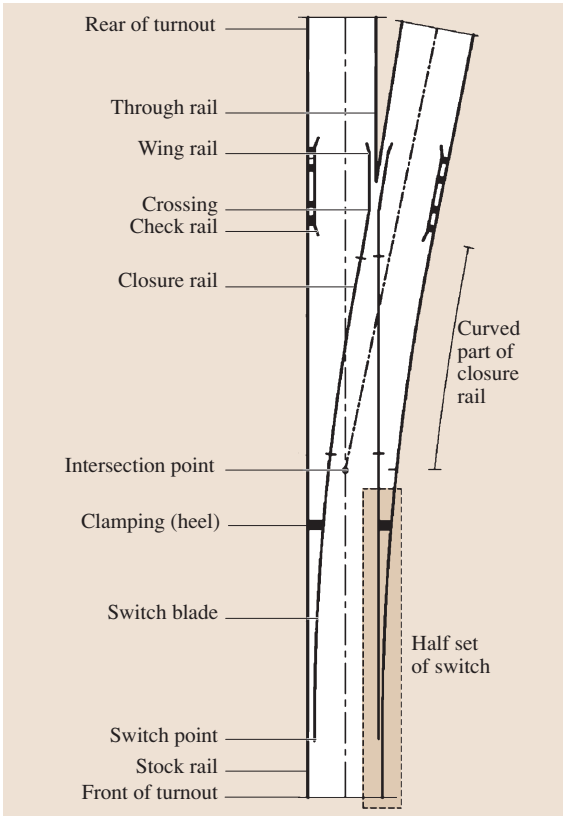


Fig. 13.107 Switches and their elements

2. The diagnosis criteria must be adaptable live during vehicle operation. This means that software skills must be available in the maintenance groups.

The reason for this is that the supplier cannot deliver a diagnosis system fulfilling all needs of the operator because:

1. Not all operation circumstances are clear for the producer or even the operator
2. Operation conditions may alter during vehicle operation (for instance, due to speed increase, longer trains, or movement from full to inferior service after 10 years or more)

It is recommended to order two or even more releases of diagnosis software after vehicle acceptance (and homologation) has been achieved. Operation experience can then be integrated and the correct level of information attained.

Helpless statements should be avoided and making sure that information is given for real problems. Also transfer from industry personnel to railway personnel must be achieved.

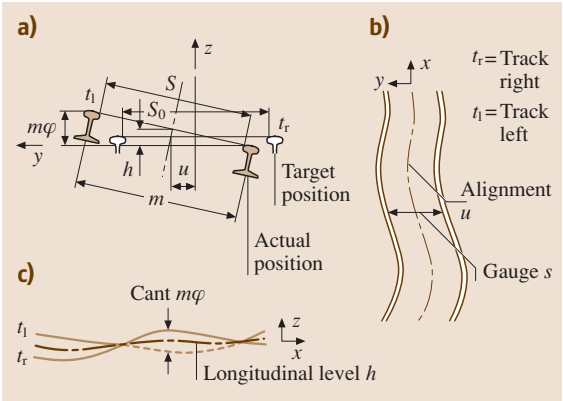


Fig. 13.108a–c Geometric track components: **(a)** coordinates in the measurement plane, **(b)** horizontal track coordinates: gauge s and alignment u, **(c)** vertical track coordinates level h and cant m_φ

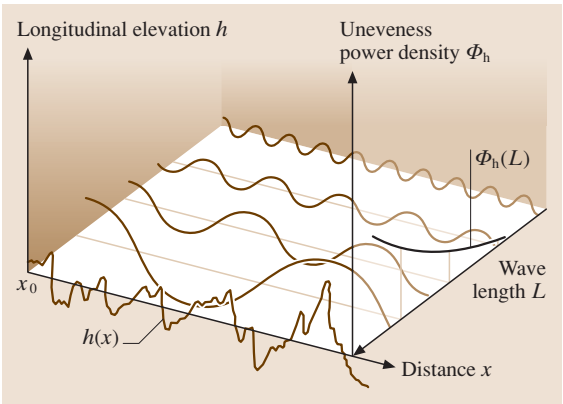


Fig. 13.109 Principle of calculation of power spectral densities (PSD) $\phi_h(L)$ from the displacement curve $h(x)$

13.3.2 Track

Interoperability of railways id affected by the track gauge very much, see Table 13.14.

Track Geometry Components

The track geometry defines the position of the track in the landscape [13.89]. The following components are used:

1. Vertical: level track, gradients, and gradient changes

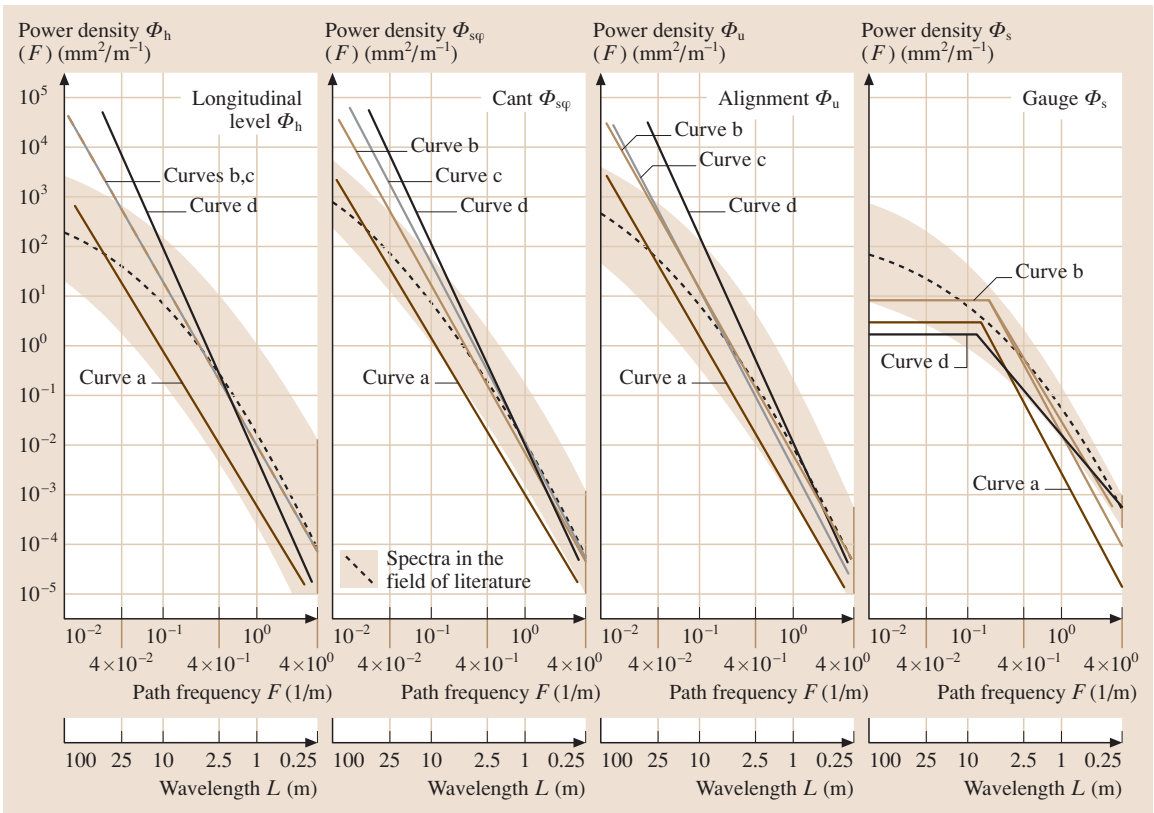


Fig. 13.110 Examples of track power spectral densities. *Curve a* – standard gauge; *curve b* – city railway flat bottom rails (route track); *curve c* – city railway flat bottom rails (driving school track); *curve d* – city railway grooved tramway rails track

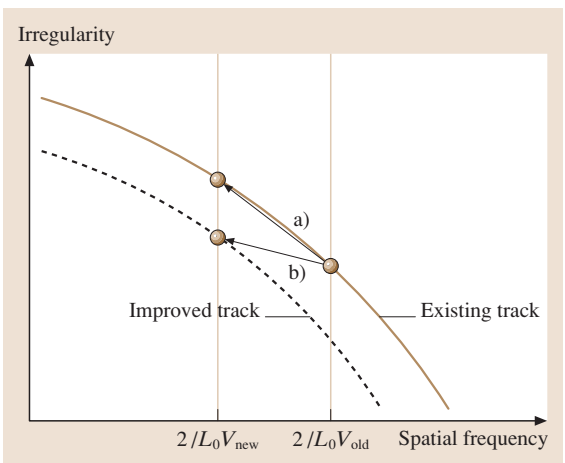


Fig. 13.111 Effect of speed increase on track amplitudes for a specific eigenfrequency of the vehicle

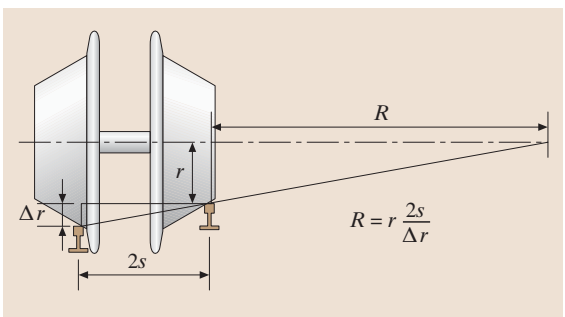


Fig. 13.112 Steering effect of a slip-free wheel-set in a tight curves (radius R) related to the inner rail. $2s$ is the distance of contact points (typically 1500 mm at standard gauge); r is the rolling radius at the inner wheel; Δr is the radius difference between the outer and inner wheel

2. Curves: radius R, transition curves with continually adapted curve radius, curvature k as inverse radius $k = 1/R$. Clothoid for curvature of transition curves $k = L/A^2$, where A is the clothoid parameter that limits the jerk r (change of acceleration) in the lateral direction (m/s^3); limit $\leq 1\,m/s^3$ (Fig. 13.100)
3. Lateral: cant to reduce lateral accelerations (maximum for standard gage with mixed traffic 150 mm) in transition curves ramps must be foreseen (Fig. 13.101).

Track Bed Configuration

The sleeper in the track bed (Fig. 13.102) not only distributes the weight in the vertical and lateral directions but also in the longitudinal direction. All forces caused by thermal stresses in the continuously welded rail must be transferred via the sleepers to the ballast. The continuous welding process is done at medium temperatures to minimize temperature effects. In summer at high temperatures buckling of the track is avoided by high lateral resistance.

Rails are described with letters, which define the shape of the rail, and a figure, which normally gives the weight per meter in kg or lb (Fig. 13.103).

The VST36 is a rail for light axle load, for instance Swiss narrow-gage railways. 49E1 is used on lightly used European standard-gage lines. 60E2 is a rail for European main lines and AREA141AB is a rail for American main lines.

60Ri2 is a grooved rail for tramway applications. Though the axle load of trams seldom exceeds 8 t the stiffness of the tram rail is similar to the other rails. In this way much smaller displacements of the rail are achieved in the street plane.

Rails today are rolled as long as possible, typically 120 m long, in order to reduce the number of weld spots in the line as much as possible. Welding in the line not only is costly but also the welds are more crack sensitive than the rolled rail.

Concrete sleepers are very durable and environmentally friendly, therefore they have replaced wooden sleepers in many cases. Their heavy weight makes assembly difficult but provides good track stability (Fig. 13.104).

Today the most commonly used sleeper type is concrete, but wood is still used also and Y-steel sleepers have a growing market-share. Y-Steel sleepers have much higher lateral resistance than concrete or wood, therefore no ballast is used to avoid sideways buckling at high temperatures. Also, because of the lower height of the sleeper, less ballast is used in the track bed itself (Fig. 13.105). This is an advantage in tunnel construction. On intensively used lines, ballast-less track construction has the advantage of less maintenance effort compared with ballasted tracks (Fig. 13.106).

Switch

To enable networks, switches are essential. Compared with other guided transport modes where switches are large and heavy, at the railway switch only the switch blades must be moved. This is done by bending the blades elastically. Relatively small electric motors with gearboxes apply these forces.

In the crossing (also called frog) itself there is a gap where no lateral guidance of the wheel flange occurs.

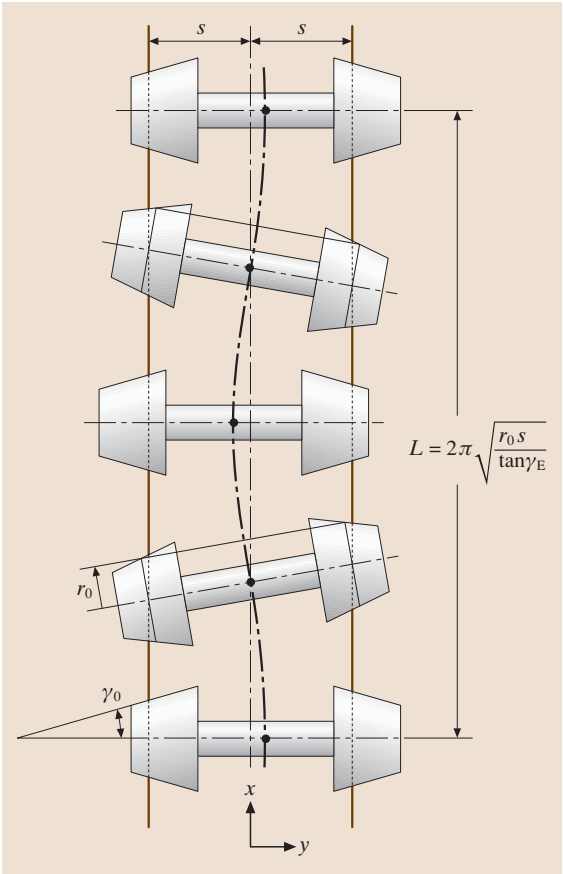


Fig. 13.113 Idealized wheel-set moving: the wavelength L which occurs in slip-free movement [13.90]

Therefore check rails and the running rail result in guidance on both sides of the other wheel of the wheelset (Fig. 13.107).

Track Irregularities

The four degrees of freedom for the track are: gauge s, cant m_φ, alignment u, and level h (Fig. 13.108).

Gauge is defined as the smallest distance between the rail heads in a track 0–14 mm under the top of rail (TOR) cant m_φ is the height difference between the two rails (Table 13.14). Twist is a function of cant over

Table 13.14 Nominal international gauge distribution (caution: large tolerances between −3 and +35 mm are possible)

Name	(mm)	British (in)	Worldwide (%)
Meter gauge	1000	$3-3\frac{3}{8}$	7.5
CAP (gauge)	1067	3–6	7.7
Standard gauge	1435	$4-8\frac{1}{2}$	64
Russian broad gauge	1524*	5	11.8

* New 1520 since about 1980

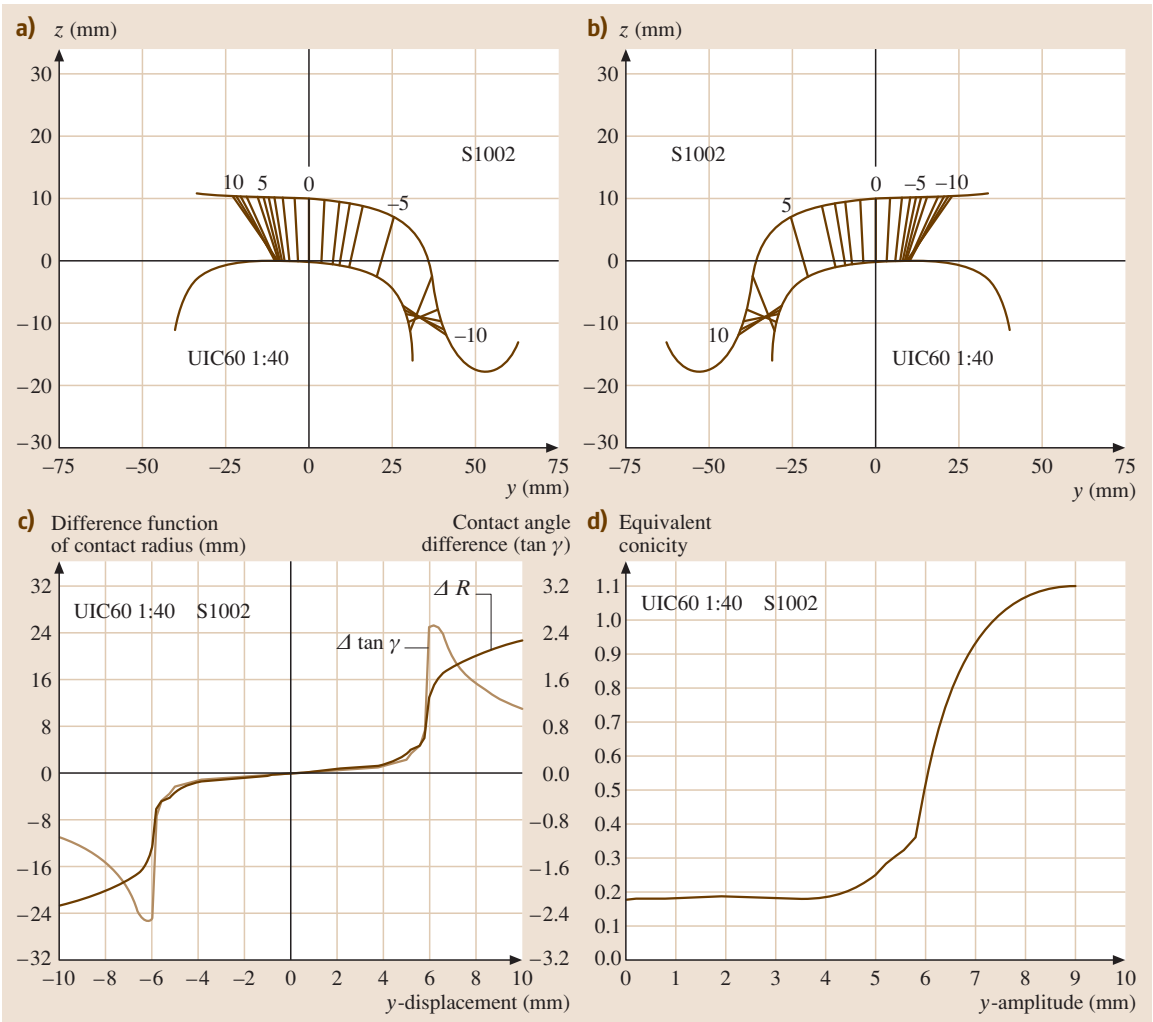


Fig. 13.114a–d Wheel–rail contact for S1002/UIC 60 1:40 inclined. Contact points and contact functions of wheel and rail, profiles wheel S1002 (wheel diameter 700 mm, flange gauge 1426 mm, wheel load 175 kN), rail UIC 60, 1:40 inclined, track gauge 1435 mm. **(a)** Left contact point, **(b)** right contact point, **(c)** difference function of contact radius ΔR and contact angle difference function $\Delta \tan \gamma$, **(d)** equivalent conicity

distance

$$\text{cant} = [m_\varphi(x_1) - m_\varphi(x_2)]/(X_2 - X_1)[-]$$

(possible also (mm/m = ‰)).

Track irregularities are defined as a function of wavelength L or spatial frequency $\Omega = 2\pi/L$. With increasing wavelength the amplitude of track irregularities increases (Fig. 13.110).

The following track irregularities are common [13.91]:

- Level

$$\Phi_z(\bar{\Omega}) = \frac{A_V \Omega_c^2}{(\bar{\Omega}^2 + \Omega_r^2)(\bar{\Omega}^2 + \Omega_c^2)} A$$

- Alignment

$$\Phi_y(\bar{\Omega}) = \frac{A_A \Omega_c^2}{(\bar{\Omega}^2 + \Omega_r^2)(\bar{\Omega}^2 + \Omega_c^2)}$$

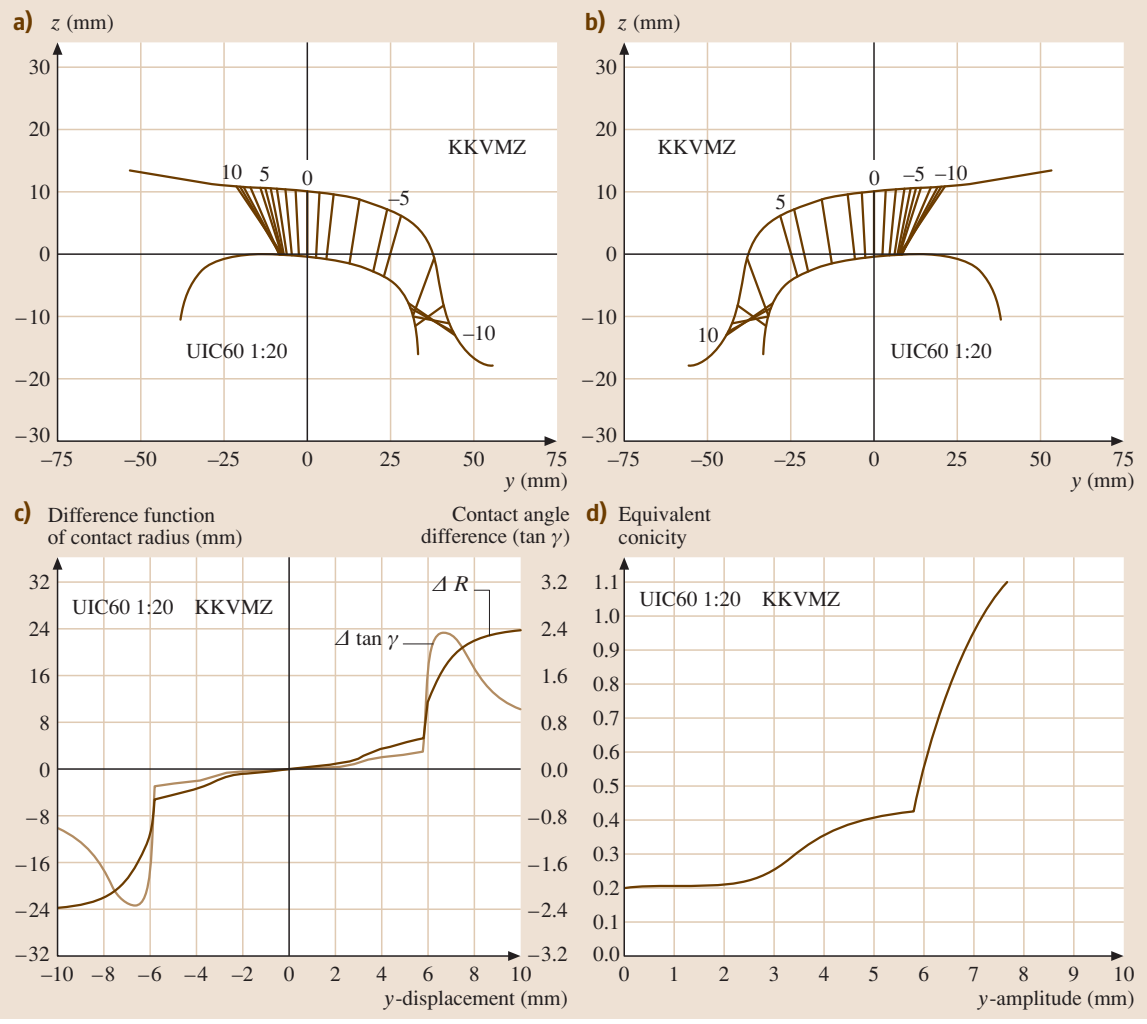


Fig. 13.115a–d Profile combination KKVMZ–UIC 60 1:20 (track gauge 1435 mm). Contact points and contact functions of wheel and rail, profiles wheel KKVMZ (wheel diameter 700 mm, flange gauge 1426 mm, wheel load 175 kN), rail UIC 60, 1:20 inclined, track gauge 1435 mm. **(a)** Left contact point, **(b)** right contact point, **(c)** difference function of contact radius ΔR and contact angle difference function $\Delta \tan \gamma$, **(d)** equivalent conicity

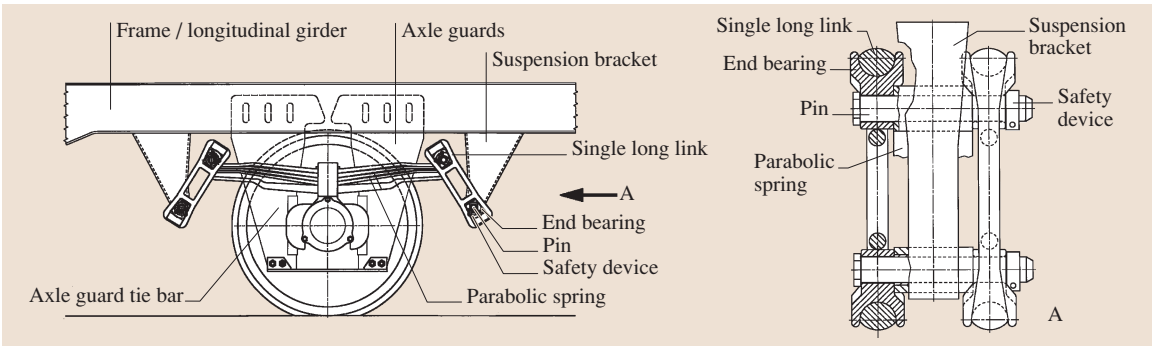


Fig. 13.116 Running gear for two-axle wagons with single long link suspension *Niesky2* (Waggonbau Niesky GmbH, Niesky)

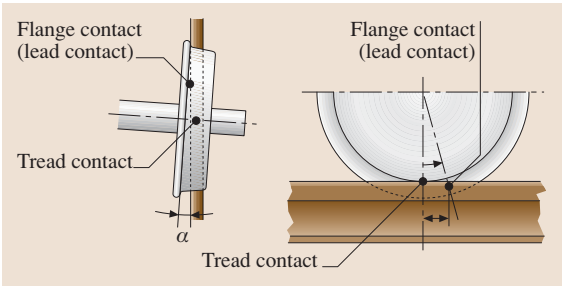


Fig. 13.117 Two points of contact in different planes, if the angle of attack between wheel and rail α exceeds a certain value. Lead contact in the flange

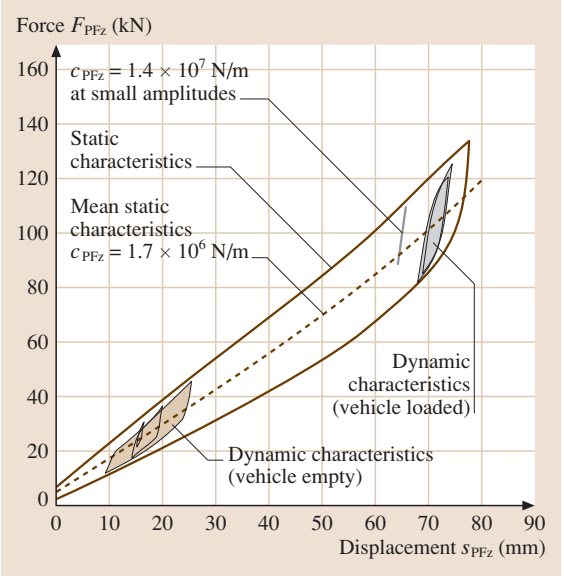


Fig. 13.118 Measured force–displacement diagram in the vertical direction of a leaf spring–link suspension system: spring rate and friction damping

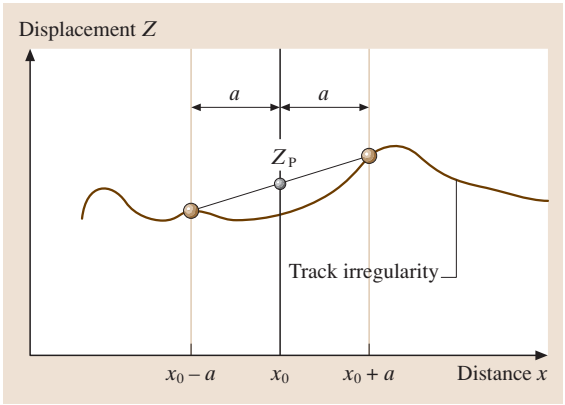


Fig. 13.119 A bogie with two wheel-sets spaced from the bogie center

- Cant in radians

$$\Phi_\rho(\Omega) = \frac{(A_C/a^2)\Omega_c^2\bar{\Omega}}{(\bar{\Omega}^2+\Omega_r^2)(\bar{\Omega}^2+\Omega_c^2)(\bar{\Omega}^2+\Omega_s^2)} ,$$

with the following data for a conventional track in good conditions: $\Omega_s = 0.4380\,\text{rad/m}$, $\Omega_c = 0.8246\,\text{rad/m}$, $\Omega_r = 0.0206\,\text{rad/m}$, $A_V = A_A = A_C = 5.9233\times10^{-7}$ m rad, $a = 0.75$ m.

Impact of the increase in speed (Fig. 13.111):

a) Without track quality improvement
b) With track quality improvement

Relevant wavelength L_0 is increasing as the eigenfrequencies f_0 of vehicles are time invariant

$$f_0 = V/L_0 .$$

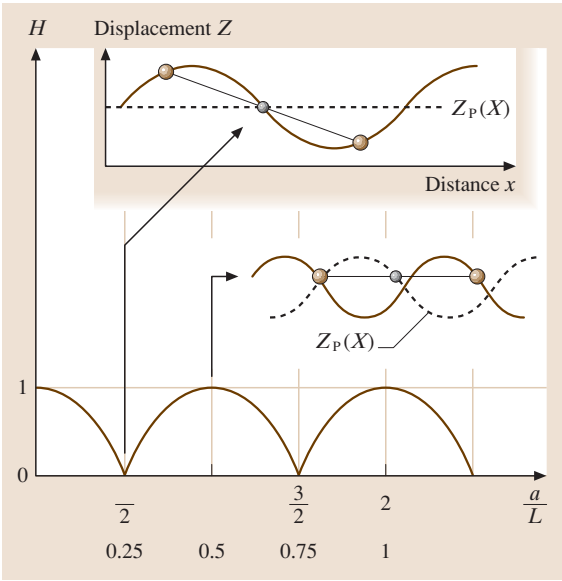


Fig. 13.120 Geometrical transfer function with values between 0 (no transfer of track irregularities) to 1 (all track irregularities are fully transferred, but not amplified)

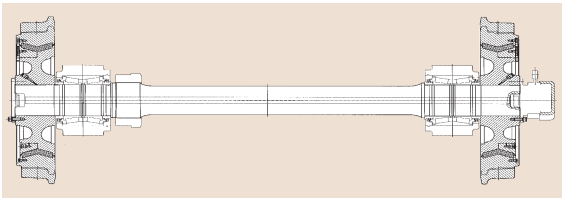


Fig. 13.121 Driven wheel-set with inside bearings and rubber elastic wheels for the low-floor tram Schwerin (Germany), wheel type B02000, running circle new 600 mm, gauge 1435 mm, weight 552 kg, without bearings, wheel stiffness: 20.0 kN/mm radial and 20 kN/mm axial (Bochumer Verein, BVV, Bochum)

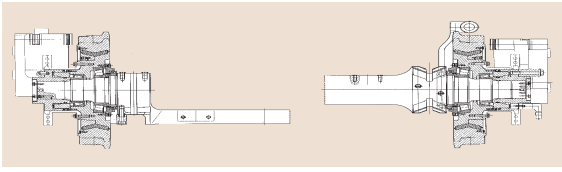


Fig. 13.122 Idle wheels with cranked axle for low-floor tram Schwerin with rubber-cushioned individual turning wheels of type BO2000, wheel diameter new 600 mm, gauge 1435 mm, weight 897 kg, with bearings and brake system, wheel stiffness: 200 kN/mm radial and 20 kN/mm axial (Bochumer Verein, BVV, Bochum)

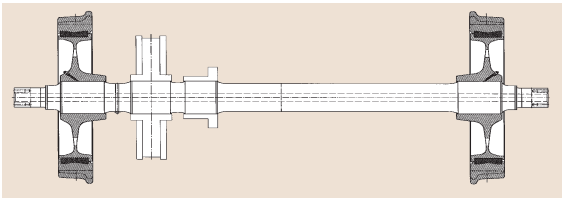


Fig. 13.123 Driven wheel-set with brake disk from the Vienna Metro, rubber sprung wheels type BO54, wheel diameter 840 mm new, gauge 1435 mm, weight 852 kg, wheel stiffness: 75 kN/mm radial and 8 kN/mm axial (Bochumer Verein, BVV, Bochum)

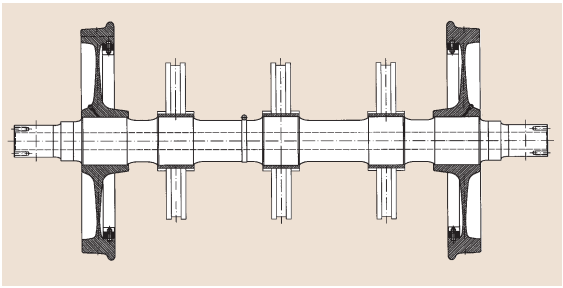


Fig. 13.124 High-speed wheel-set with seats for three brake disks and wheels with sound absorbers. The shaft is hollow bored for weight reduction and the possibility of ultrasonic crack detection, wheel diameter 920 mm new, gauge 1435 mm, weight 948 kg without brake discs (Bochumer Verein, BVV, Bochum)

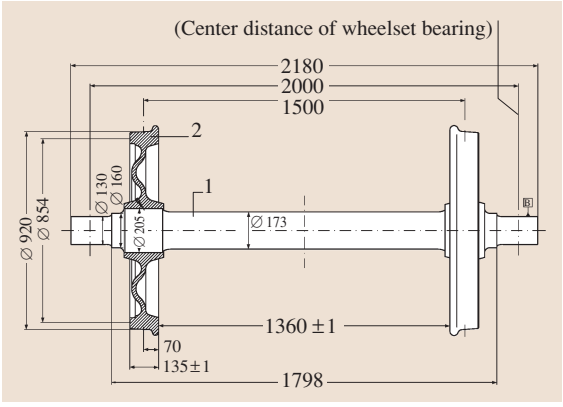


Fig. 13.125 Standard freight car wheelset type BA 304 for 25 t axle load. 1 – shaft, 2 – wheel with bell shaped web to reduce stresses from block braking, wheel diameter 920 mm new, 854 mm worn, weight 1003 kg (Radsatzfabrik Ilsenburg Rafil)

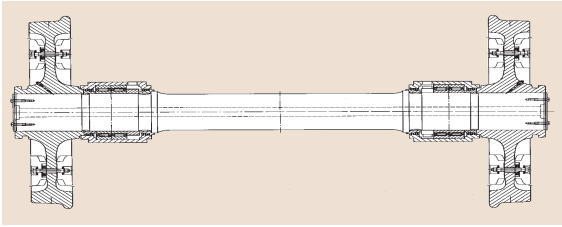


Fig. 13.126 Leila freight car wheel-set for 22.5 t axle load with inside bearings and wheel brake-disks 920 mm wheel diameter new, gauge 1435 mm, weight 1392 kg with bearings and aluminum brake discs◀

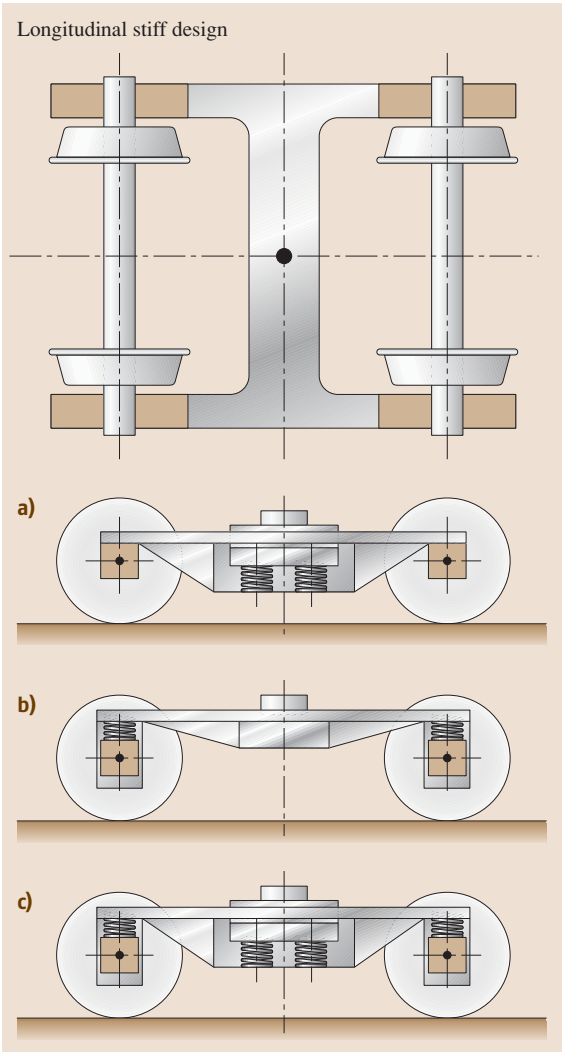


Fig. 13.127a–c Basic concepts of bogie suspension: **(a)** only secondary sprung (three-piece bogie), **(b)** only primary suspension (union international des chemin de fer (UIC) freight bogie), **(c)** with primary and secondary suspension; all in longitudinal stiff design

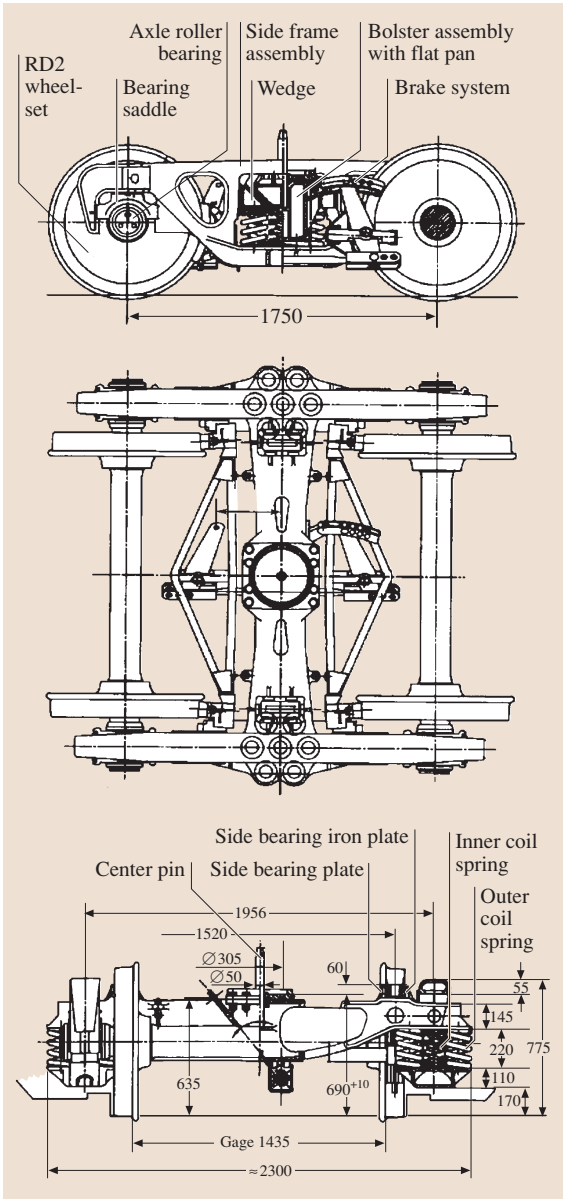


Fig. 13.128 Z8A-Freight car bogie (China) weight 4100 kg, max. axle load 210 kN (Qiqihar railway rolling stock Co., China) (RD2 is a wheelset type name)

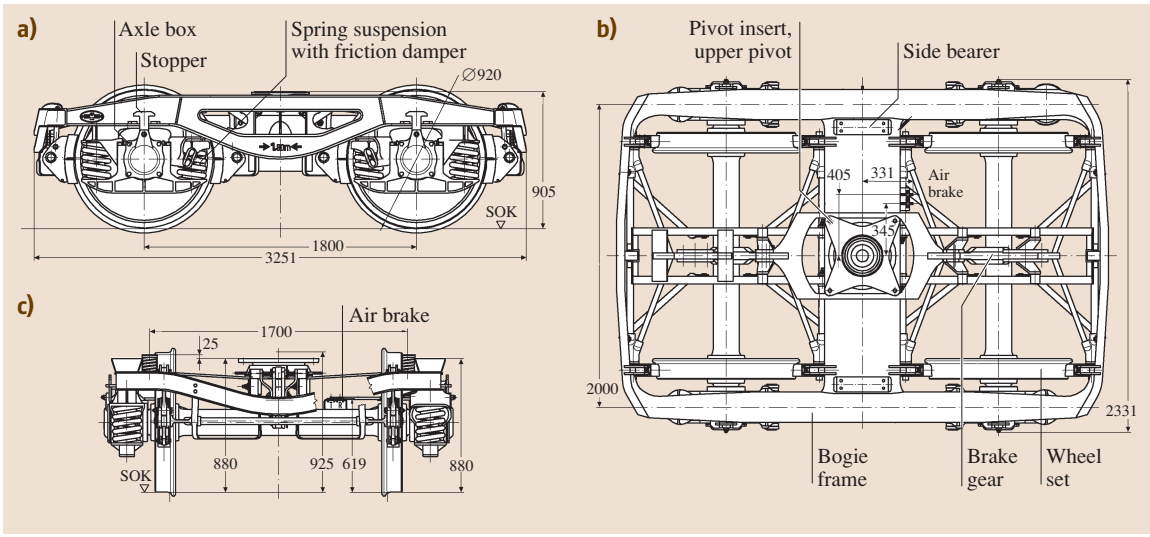


Fig. 13.129 Freight bogie Y25Lsd1-K with K-brake block insert in single brake block arrangement for axle load 22.5 t, mass 4390 kg (with wheel-sets type BA 004 and pivot) (Eisenbahn Laufwerke Halle, ELH)

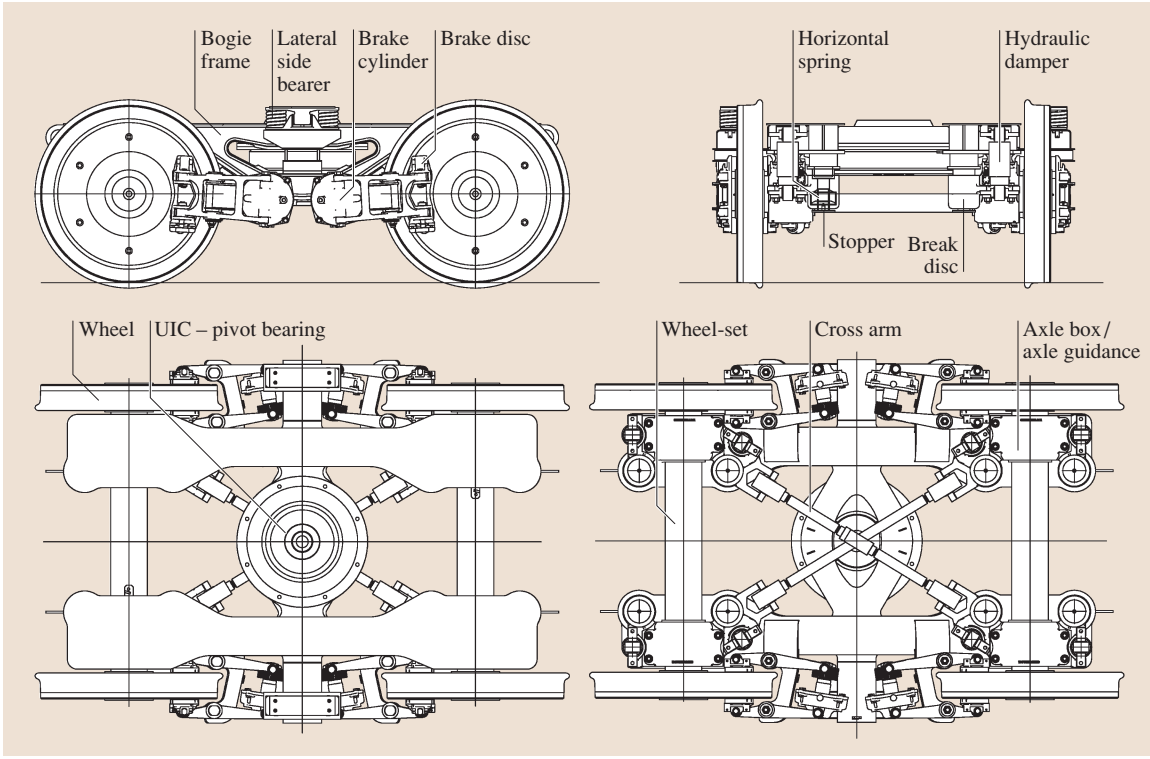


Fig. 13.130 Leila freight bogie (Josef Meyer Waggon AG, Rheinfelden)

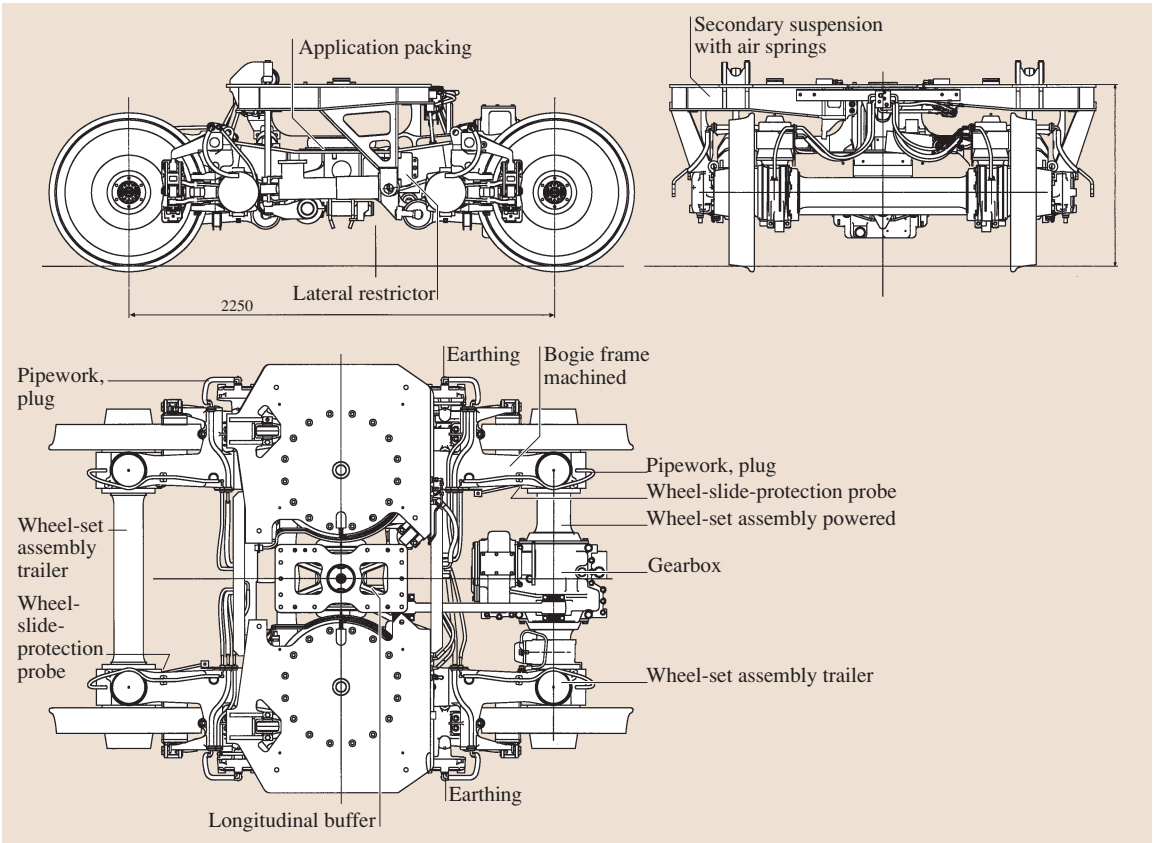


Fig. 13.131 Bogie assembly type A type B5000 for axle load 16 t, weight 4700 kg powered, wheel diameter new 780 mm, worn 716 mm, wheelbase 2250 mm, V_{max} 200 km/h (Bombardier Transportation, Berlin)

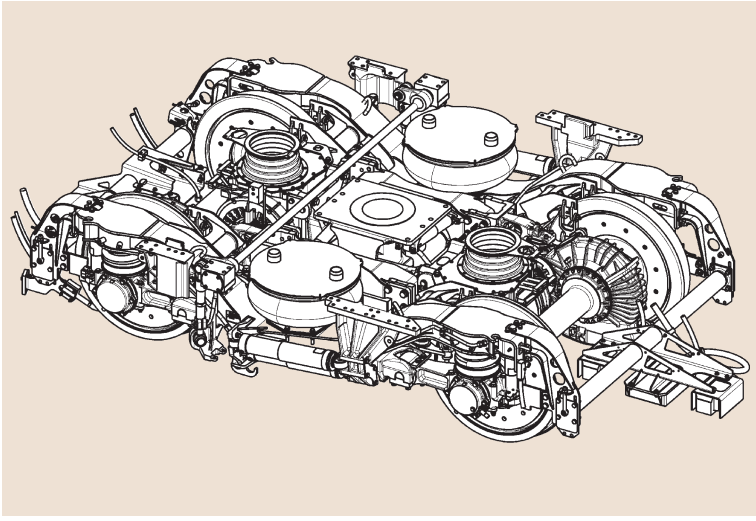


Fig. 13.132 Running gear SF 500-type motor bogie for operating speed up to 350 km/h, wheel-set load maximum 17 t, continuous power per wheel-set up to 500 kW, maximal starting tractive effort per wheel-set 19 kN, wheel-set distance 2500 mm, gauge 1435 mm, wheel diameter new/worn 920/830 mm minimal curve radius service/workshop 150/120 m, weight with pivot and traverse 9.2 t (Siemens Transportation Systems, Erlangen)

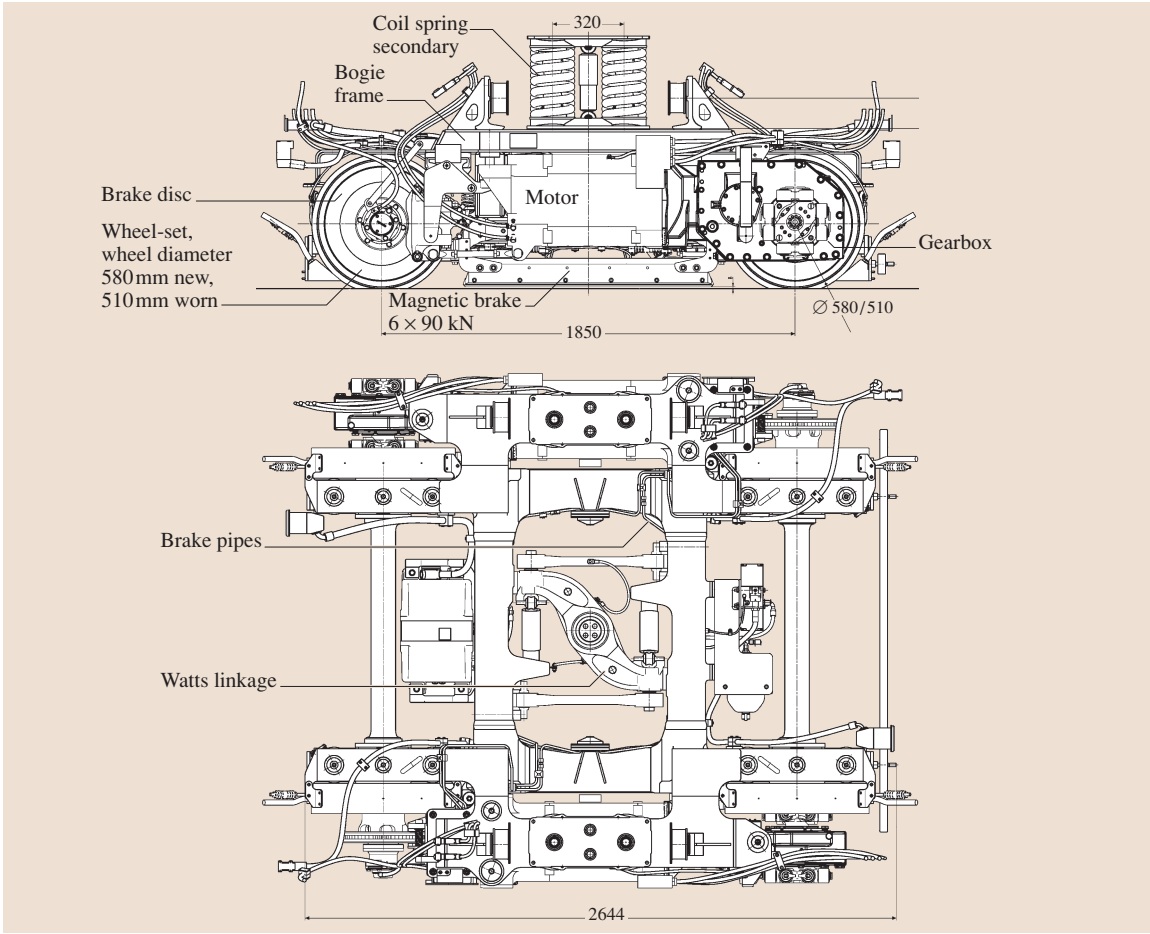


Fig. 13.133 Bogie assembly type S1000 BM1000 tram Marseille MB, weight 4700 kg powered, maximum speed 70 km/h (Bombardier Transportation, Berlin)

13.3.3 Running Gears

Wheel–Rail Interaction

The standard element of railway running gears is a wheelset. This consists of a rotating shaft and two wheels which are fixed on this shaft by a press or shrink fit.

Additionally the wheel profile together with the rail profile generates a steering effect (Fig. 13.112).

To introduce this effect a conical thread geometry is first assumed.

The radius difference Δr in both rolling wheels leads to a self-steering effect, as both wheels have the same rotational speed but the outer wheel is running on a larger radius than the inner wheel.

This desirable behavior is superimposed by an effect called sinus running for conical profiles or wave running for practical profiles.

Equivalent Conicity. The wavelength of the real profile is equal to the wavelength of a wheel profile with constant conicity, as shown in Fig. 13.113.

To characterize the interaction between wheel and rail for different wheel and rail profile shapes Figs. 13.114 and 13.115 show examples for two different rail inclinations. The desired behavior is that the surfaces of wheel and rail make contact on a rather broad level and that the difference in radius Δr reaches high values before flange contact occurs. If the angle of attack between wheel and rail is large, then a two point contact occurs (Fig. 13.117).

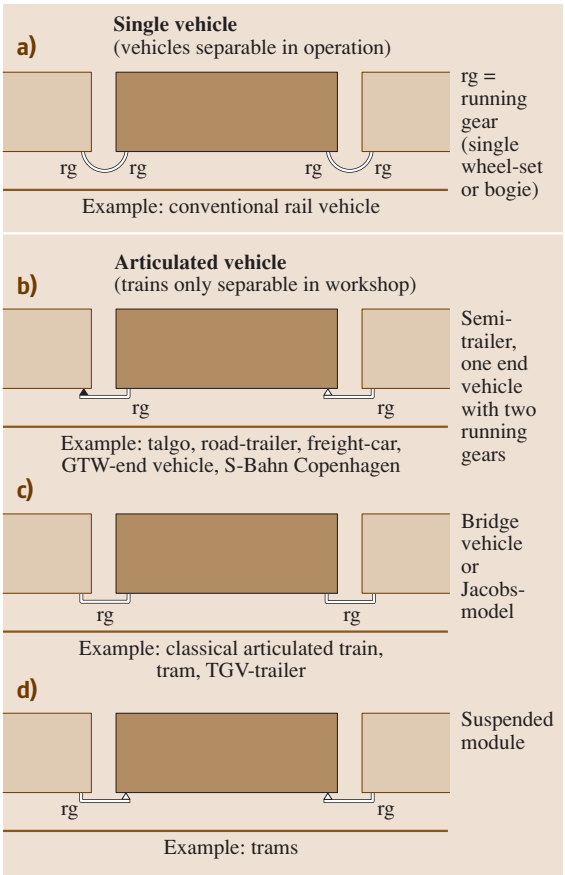


Fig. 13.134 **(a)** Single vehicle supported by at least two running gears (single wheel-sets or bogies). **(b–d)** Articulated vehicles: saddle principle, Jacobs principle, bridge principle, and suspended modules (GTW = Gelenktriebwagen (German) = articulated vehicle)

The running gears of the two-axle wagon are of the link suspension type (Fig. 13.116). The links not only admit lateral but also longitudinal movements. In this way self-steering of the individual wheelsets is enabled. The system is damped by friction, which is load sensitive; the higher the load the greater the friction force. The diameter of the link is increased in the contact zones to increase the friction force.

The spring rate is very much a function of the amplitude. For small amplitudes, for instance, caused by small track irregularities, the system is very stiff. For large amplitudes, for instance, caused by severe rail twist, the system becomes rather soft [13.92] (Fig. 13.118).

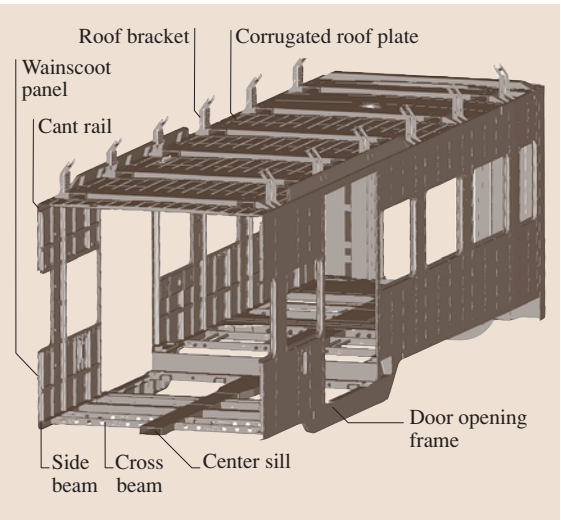


Fig. 13.135 Sheet and stringer design in steel (FTD Fahrzeugtechnik Dessau AG)

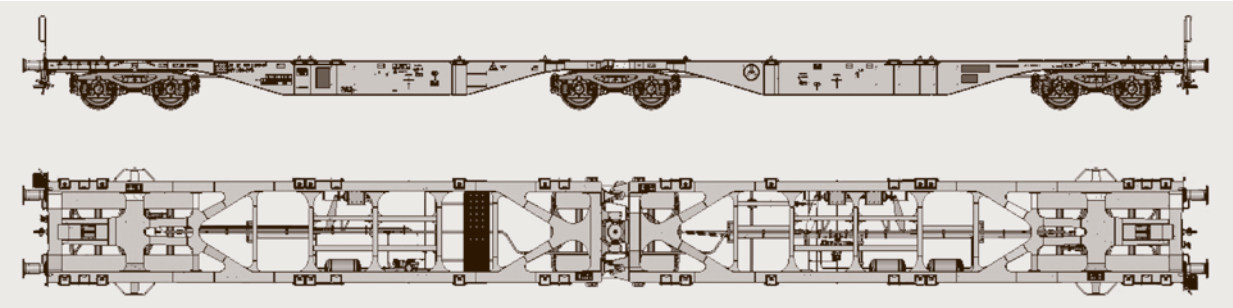


Fig. 13.136 Six-axle, three-bogie articulated flat bodied unit Sggmrss-90 for container and swap bodies: axle load 22.5 t, load height 1155 mm, length over buffers 29 590 mm, empty weight 27.6 t, V_{max} 120 km/h (with 20 t axle load), bogie type Y25Ls(s)d1 (Trinity Rail Europe)

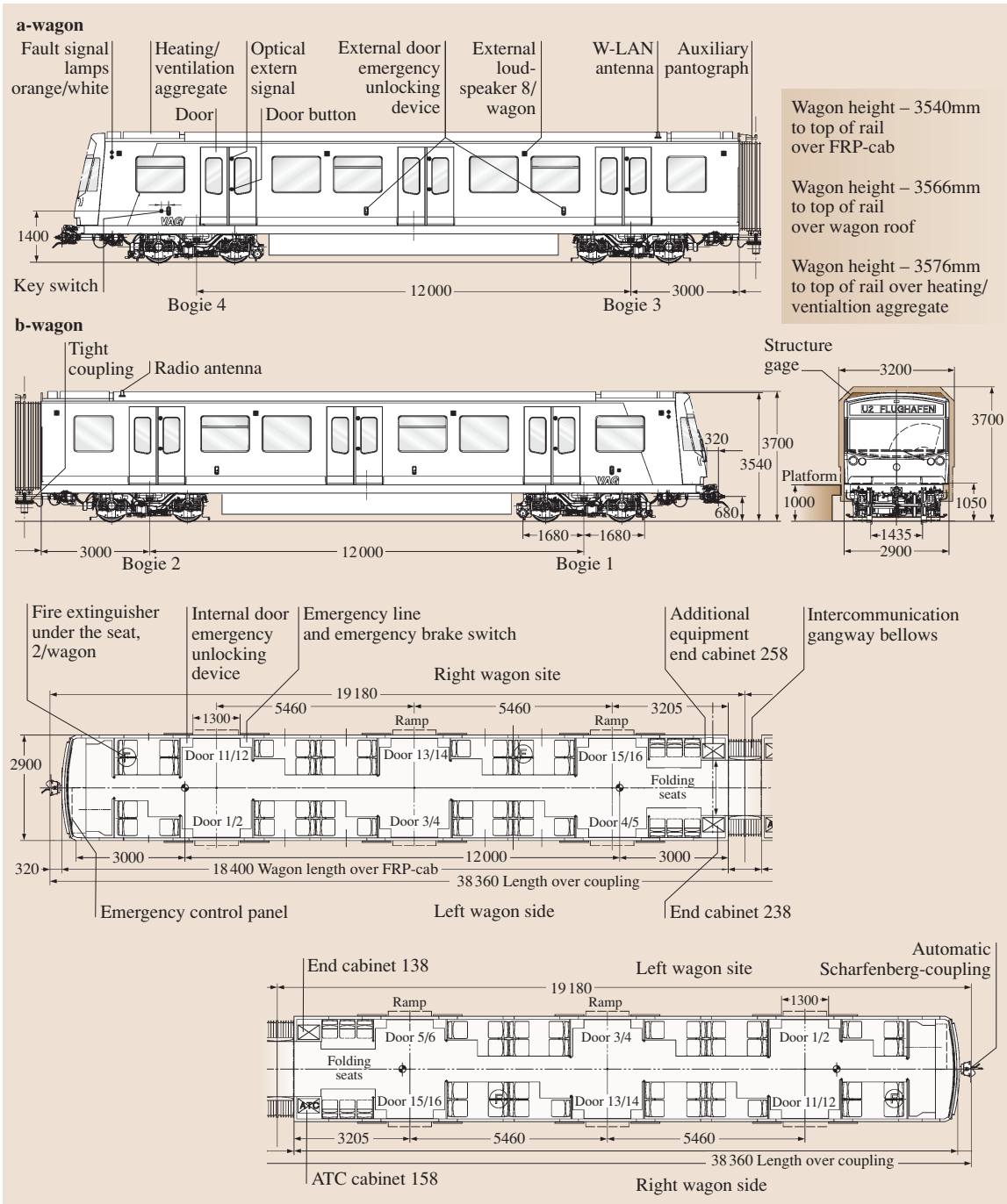


Fig. 13.137 DT3 RUBIN (automatic, driverless metro system) Nürnberg with auxiliary driver desks, but no cabs. Body-shell aluminum extrusion profiles, empty weight 59.2 t, loaded weight 98.4 t, 82 seats, capacity 424 passengers (6 /m^2) (Siemens Transportation Systems, Erlangen) (ATC = automatic train control, FRP = fiber-reinforced plastic)

Bogie Principle

The wheelsets are spaced with an axial distance $2a$ and situated in a frame (Fig. 13.119). This frame together with the wheelsets forms the bogie. The vehicle body in general is supported in the middle of the bogie frame at z_P. This geometrical configuration leads to the reduction of track regularities. As a function of their wavelength, track irregularities are reduced between two extremes: no reduction at all (for wavelengths equal to the wheel set distance $2a$) and complete reduction (for wavelengths equal to half of the wheelset distance), as illustrated by the equations below:

Geometrical transfer function of a bogie

$$z_P(x_0) = \left[z(x_0 - a) + z(x_0 + a)\right]\frac{1}{2}.$$

Track irregularity $z(x)$

$$z_P(x_0) = \frac{1}{2}\left(z\,e^{i\Omega(x_0+a)} + z\,e^{i\Omega(x_0+a)}\right)$$

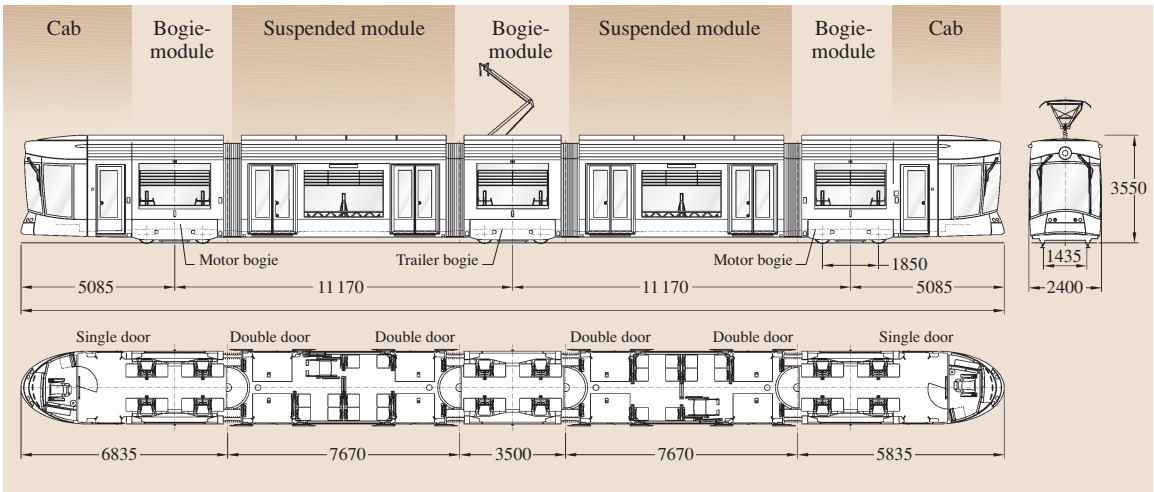


Fig. 13.138 Type of vehicle: Bombardier Flexity outlook, model Marseille: bidirectional, length 32.5 m, height 3.5 m, width 2.4 m, floor height above top of rail (low floor entrance) 320 mm, wheel diameter new/worn 580/520 mm, gage 1435 mm, car weight (empty) 40 t, car weight (loaded) (4 pass./m^2) 54.3 t, maximum axle load 11 t, minimum horizontal curve radius 25 m, minuimum vertical curve radius, crest 450 m, minimum vertical curve radius, sag 350 m, maximum speed 70 km/h, maximum gradient 80‰, nominal current supply: 750 V$_=$, regeneration of energy, low voltage: 24 V$_=$, four three-phase asynchronous motors, motor power: 4 × 115 kW, air-cooled motor, two powered bogies/one trailer bogie, rubber/metal primary suspension, coil spring secondary suspension, eight sanders, anti-slip, anti-skid system, electrical service brake: regenerative, mechanical service brake: disc brake, magnetic brake: 6 × 90 kN (Bombardier Transportation)

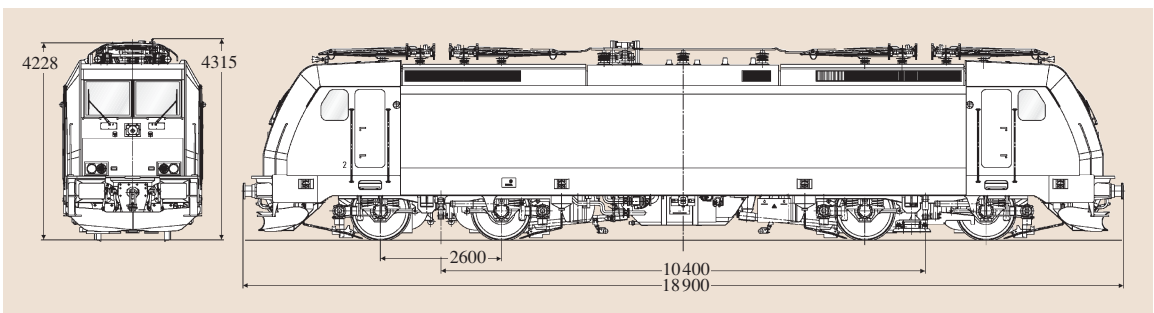


Fig. 13.139 Bombardier Traxx locomotive family, freight and passenger locomotive: weight 86 t, power 5600 kW, max. traction force 300 kN, speed 140 or 160 or 200 km/h

$$z(x) = z\,e^{i\Omega x} \qquad \Omega = \frac{2\pi}{L}$$
$$z_e^{i\Omega(x_0-a)} = z_e^{i\Omega x_0}\,e^{-i\Omega a}$$
$$z_P(x_0) = \frac{1}{2} z\,e^{i\Omega x_0}\left(e^{-i\Omega a} + e^{i\Omega a}\right)$$
$$H = \frac{\text{Output}}{\text{Input}}\,,$$
$$H = \frac{1}{2}\left(e^{-i\Omega a} + e^{i\Omega a}\right)\,,$$
$$e^{i\Omega a} = \cos\Omega a + i\sin\Omega a\,,$$
$$H = \frac{1}{2}\left(\cos\Omega a\, i\sin\Omega a + \cos\Omega a + i\sin\Omega a\right)$$
$$= \cos\Omega a$$
$$= \cos\frac{2\pi a}{L}\,.$$

For railways the lateral dynamics very often are even more important than vertical dynamics and this good behavior of a bogie is valid for the lateral direction also (Fig. 13.120).

Constructive Elements

Wheelsets. For tramways, where the track is always very stiff and therefore rather soft, rubber-cushioned wheels must be used. They also reduce noise (Fig. 13.121).

Low-floor trams do not enable wheelsets but instead need cranked axles, so that the height of the floor can be reduced (Fig. 13.122).

Many wheelsets are hollow-bored. This reduces weight and also provides the ability for ultrasonic testing (Figs. 13.123, 13.124, 13.126).

Bogies

Figure 13.128 shows the so-called three-piece bogie. This is the most common freight bogie type in the world. Several Mio bogies of this type are running outside Europe. There is only a secondary spring according to Fig. 13.127a. The three pieces that give the name to the bogie are the two side frames and the bolster assembly. The wedge (Fig. 13.128) applies load-sensitive damping.

The three-piece bogie has unsuspended side frames, whereas the Y 25 bogie (Fig. 13.129) has individual suspended axle boxes, so-called primary suspension, according to Fig. 13.127b.

The bogie type mainly used in Europe is the so-called Y 25 (Fig. 13.129). The helical springs are responsible for vertical and lateral suspension. Over inclined links in Fig. 13.127a longitudinal friction force is caused in the axle guides which damps the vertical and the lateral movements. This force is load related.

The Leila freight bogie (Fig. 13.130) enables better load distribution by internal bearings and radial steering by the cross arm. The wheelsets are those from Fig. 13.126.

In passenger transport as well interior bearings offer huge benefits with about 30% less bogie weight (Fig. 13.131).

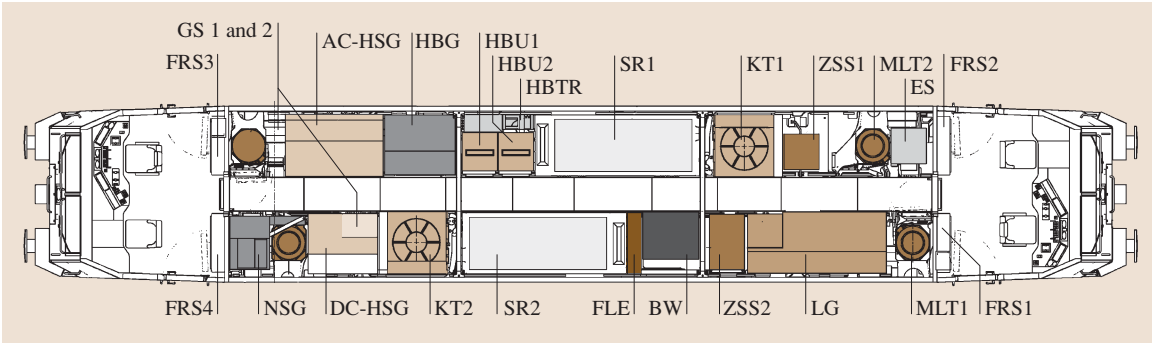


Fig. 13.140 Bombardier transportation (GS 1 and 2 = various nonpermanent equipment (inventory) like track shoes, earthing rod gloves, tools, etc.; AC-HSG = high voltage cell AC; AC = alternative current; HBG = auxiliary power distribution and battery charger; HBU1 and 2 = auxiliary converter; HBTR = auxiliary energy supply transformer; SR1 and 2 = traction converter; KT1 and 2 = cooling tower with blower; ZSS1 and 2 = signal equipment (automatic train control); MLT1 and 2 = traction motor blower; ES = electronic control equipment; NSG, NSGAT = low voltage DC distribution; DC = direct current; DC-HSG high voltage cell DC; FLE = fire detection and extinguishing equipment; BW = braking resistor; LG = compressed air supply and braking equipment; FRS1, 2, 3, 4 = cab rear wall cabinets)

The driving motor is located in the body shell and the momentum is transferred to the axle-hung gearbox via a cardan shaft (Fig. 13.131).

Figure 13.131 shows a bogie for high speed tilting trains with internal bearing wheelsets. With this assembly one obtains huge benefits in terms of weight reduction and good access to the wheels.

The bogie shown in Fig. 13.132 is a very-high-speed bogie for the fastest scheduled passenger service of today.

Low floor trams need bogies with free space in the center (Fig. 13.133).

13.3.4 Superstructures

Principle

With articulated vehicles (Fig. 13.134) normally fewer wheelsets than with standard designs are needed. Fewer wheelsets per car length means less weight, less cost, and less noise, but the maintenance process is more complicated.

The body shells are made either from aluminum intrusion profiles [13.93] or from weldable steel in a sheet and stringer design. This design offers more possibilities to adapt the structure to

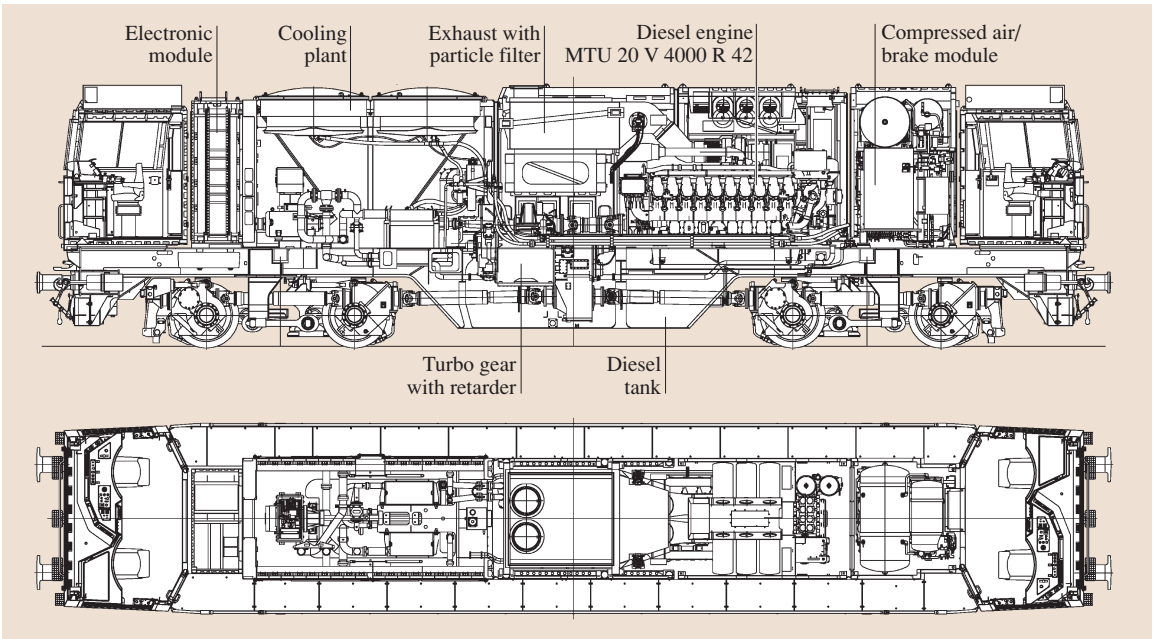


Fig. 13.141 Diesel locomotive MaK 2000-4 BB, axle arrangement BB, track gauge 1435 mm, weight 90 t, length 17 400 mm, height 4259 mm, width 3080 mm, wheel diameter new/worn 1000/920 mm, maximum speed 120 km/h, minimum radius of curve 80 m, starting traction effort 292 kN, diesel engine performance 2700 kW (Vossloh Locomotives GmbH)

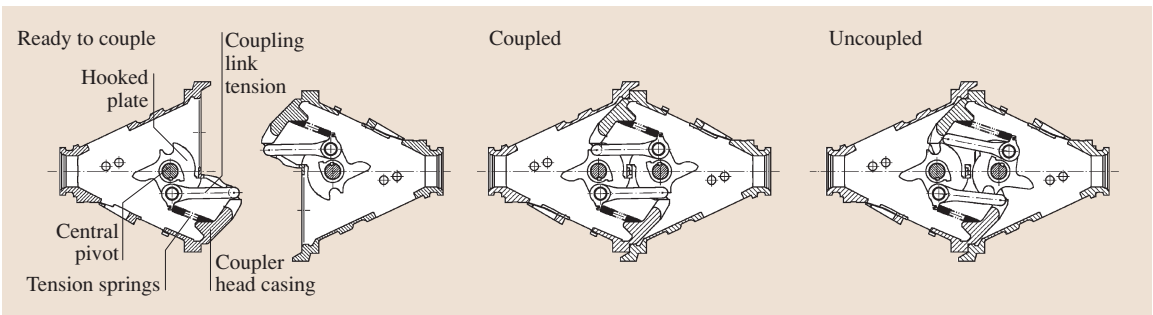
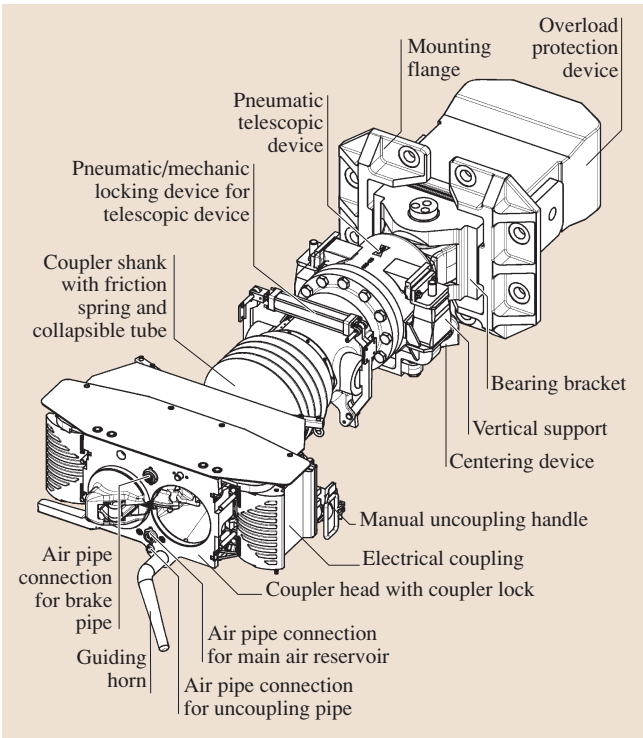


Fig. 13.142 Scharfenberg coupler, working principle (Voith Turbo)



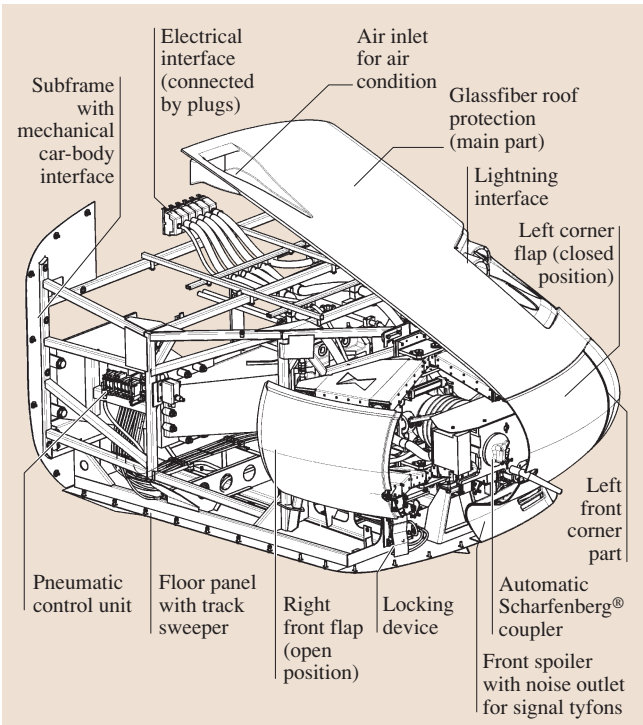


Fig. 13.143 Scharfenberg coupler for high-speed trains ICE 3 (ET-415/417)/Velaro S (AVE S 103)/Talgo 350 (AVE S 102) (Voith Turbo)◀

the load requirements and enables simpler repairs (Fig. 13.135).

13.3.5 Vehicles

Figure 13.136 shows a freight wagon as an example of a very lightweight construction. The empty wagon weighs only 20% of the fully loaded vehicle. For both situations, loaded and empty, the safety requirements must be fulfilled.

To reduce operation expenses modern metro systems are planned as driverless systems. In this way no space is lost for the driver cabin and all the space may be used by passengers (Fig. 13.137).

Modern tram cars are being built in a modular manner so that capacity can be adapted to local needs. This means that the tram shown in Fig. 13.138 can be lengthened or shortened by further bogie and suspended modules.

Electric locomotives are more powerful than diesel locomotives. To enable free running across borders they are usually equipped to cover all four electric systems in use in Europe (Figs. 13.139 and 13.140). Modern locomotives also fulfill crash concepts, see the section on *Passive Safety* and [13.94].

Diesel locomotives may either have electric or hydrodynamic power transmission. Figure 13.141 shows an example of hydrodynamic transmission.

13.3.6 Coupling Systems

As trains are formed from several vehicles the coupling device between the vehicles is essential. It transmits not only high forces, but also data channels. Coupling and uncoupling must be done very reliable (Figs. 13.142 and 13.143).

Fig. 13.144 Scharfenberg front nose for high-speed trains Talgo 350 (AVE S 102) (Voith Turbo). Technical data: overall length 2985 mm, overall hight 2042 mm, overall width 2866 mm, weight 1020 kg (without coupler), maximum train speed > 350 km/h, aerodynamic loads ± 11 kPa, mechanical carbody interface 10 screws M 16, frontflaps opening angle 64°, material of outside parts, GRP (glassfiber reinforced plastic) sandwich laminate, fire protection according to DIN 5510 S4◀

Coupling

The cone-and-funnel shape of the coupler front face profile ensures a generous gathering range both horizontally and vertically and allows automatic coupling on curves, even with vertical mismatch and very low speed. Minimal force is required for successful coupling.

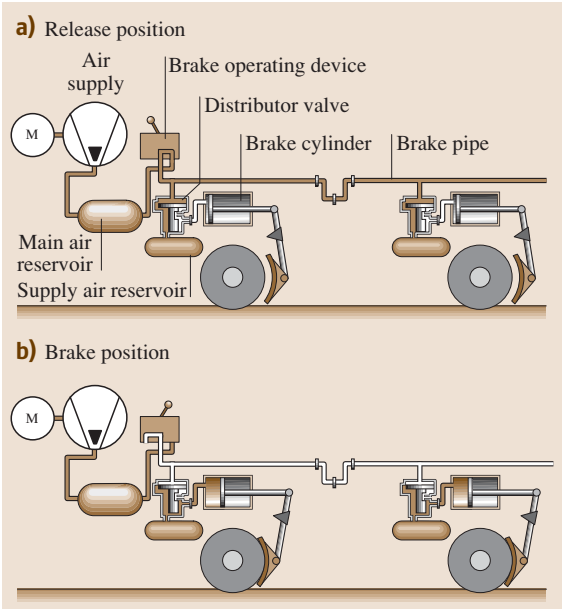


Fig. 13.145a,b Indirect (automatic) train brake. **(a)** Release position, **(b)** brake position

Coupled

The coupler faces and the locking system form a rigid connection both vertically and horizontally. The parallelogram arrangement of the coupler link provides uniform distribution of the draft load. This coupler link design ensures minimal wear and maximum coupler longlivity. The rigid and slack-free connection enables jerk-free acceleration and braking, and offers optimum ride comfort. It also prevents the cars overriding one another in the case of an accident.

Uncoupling

The geometry of the coupler lock enables automatic uncoupling even under traction load. Automatic uncoupling is ensured when misaligned, even on gradient changes. The uncoupling operation is irreversible. SCHAKU's safety philosophy does not allow recoupling, unless the cars have moved apart and the couplers have separated.

13.3.7 Safety

Active Safety Systems/Brakes

Brakes transform the kinetic energy of a train into other forms of energy. There are three principle tasks for a brake: bringing the train to a halt (stop braking), maintaining the speed of the train on a gradient (downhill braking), and preventing a stationary train from moving (park braking).

From the mechanical point of view brakes can be categorized into adhesion-dependent and non-adhesion-

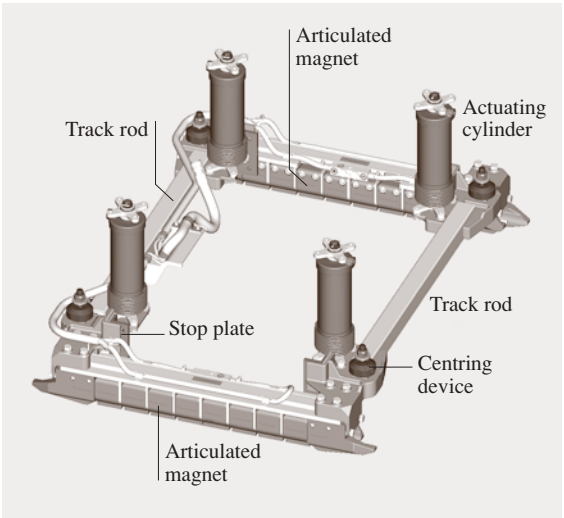


Fig. 13.146 Track brake, speed up to 200 km/h

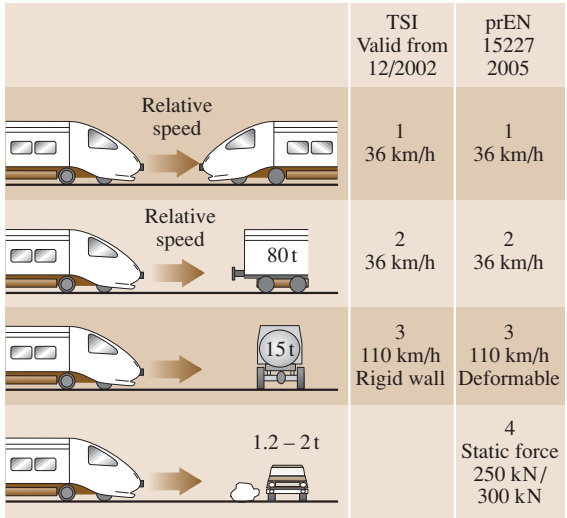


Fig. 13.147 Crash scenarios

dependent brakes. The former type always work via brake moments to wheels, whereas the latter are track brakes or aerodynamic brakes. Adhesion-dependent brakes may be friction brakes (tread or disc brakes) or dynamic brakes (electrodynamic brakes, where a motor works as a generator, or hydrodynamic brakes, which have a hydraulic retarder).

For safety reasons all rail vehicles must be equipped with an indirectly acting pneumatic brake and the brake pipe must run through the entire train (Fig. 13.145).

The indirectly acting brake operates according to the following principle. If the full brake pressure (typically 5 bar) is available and the brake pipe is on that pressure, then the same pressure is in the supply air reservoir and the brake cylinder is released by a mechanical spring in the brake cylinder. If the brake pressure is reduced, for instance, by the brake operating device, then the distributor valve connects the supply air reservoir with the brake cylinder and the brake is applied.

The brake is released by increasing the brake pipe pressure again. Then the distributor valve releases the pressure from the brake cylinder into the atmosphere and simultaneously refills the supply air reservoir with the brake pipe pressure. Because of this indirect or automatic behavior the brakes are also applied if the train breakes apart and the pipe is separated.

Also the brake can be applied from any location along the brake pipe, for instance, by a control van or by an emergency brake device.

The disadvantage of an indirect brake is that it takes a long time for application (up to 30 s for freight trains) and release (up to 60 s for freight trains). Therefore vehicles that must be controlled quickly and precisely, for instance, locomotives, must be equipped with an additional direct brake device.

The brake actuators may operate through one-sided (Fig. 13.128) or two-sided (Fig. 13.129) brake blocks. Because of wear and noise demands, modern brake blocks are no longer made from cast iron but from composite materials.

Higher thermal capacity can be achieved by disk brakes: either wheel disk brakes (Figs. 13.130 or 13.131) or shaft disc brakes (Fig. 13.124).

To reduce the braking distance further rail brakes can be used as they are not dependent on the wheel–rail friction coefficient (Fig. 13.146). If the brake is not used pneumatic cylinders lift the brake to avoid contact because of track irregularities. The magnets are excited by direct current from batteries. The friction force between the magnet and the rail because of the magnetic forces cause the brake force.

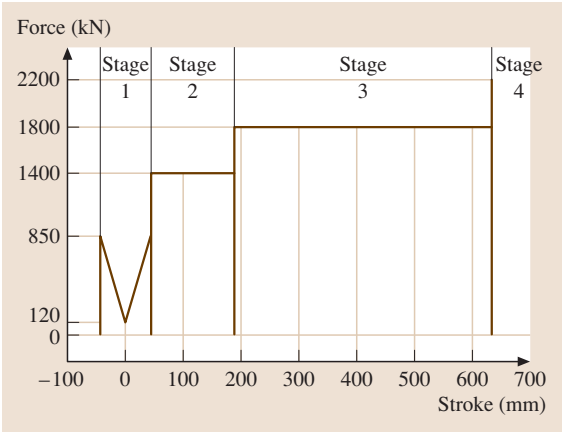


Fig. 13.148 Energy absorption of the coupler from Fig. 13.143. Overall length 2345 mm, telescopic stroke 200 mm, weight 960 kg, vertical swing ±6°, horizontal swing ±20°, tensile strength 850 kN (braking strength 1000 kN), compressive strength 1400 kN (breaking strength 2000 kN), minimum coupling speed 0.6 km/h, admissible impact speed allowing buffer to regenerate: 5 km/h, admissible impact speed before coupler tear-off: 20 km/h. Energy absorption capacity. Stage 1: friction spring 120–850 kN, 44 mm stroke; Stage 2: collapsible tube 1400 kN, 145 mm stroke; Stage 3: collapsible tube 1800 kN, 445 mm stroke; Stage 4: shear-off elements 2200 kN; total energy absorption 1025 kJ

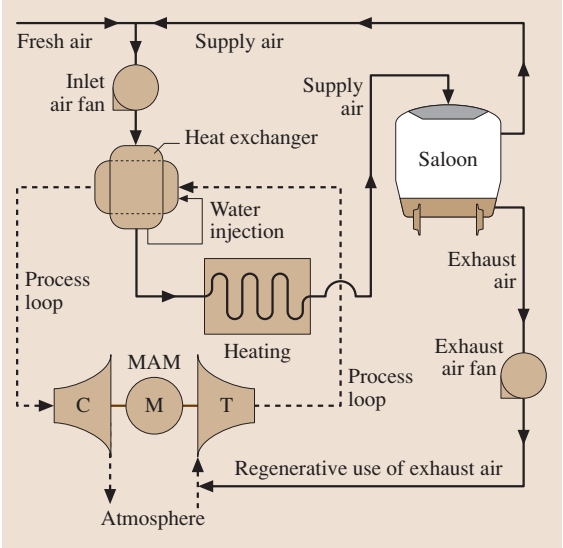


Fig. 13.149 Air cycle concept (Liebherr) (C = compressor; M = motor, T = turbine, MAM = motorized air cycle machine)

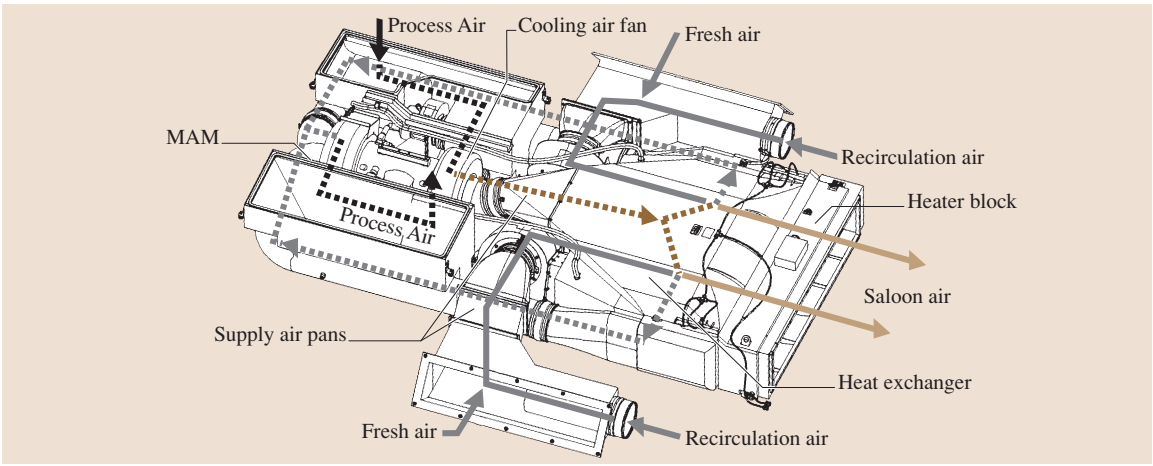


Fig. 13.150 Components and air flow (Liebherr). Roof-mounted air-conditioning unit for high-speed trains, weight 550 kg

For trams the lifting device can be avoided because of the lower speeds (Fig. 13.133).

Passive Safety

Though the active safety of railway systems is very high due to the implementation of signals, brakes, and train control systems [13.95] collisions cannot be totally avoided. Therefore it makes sense to reduce the potential human factor through the use of safety devices.

For various crash scenarios (Fig. 13.147) one must ensure that no severe injuries occur to passengers and staff.

Crash Scenario:

- Scenario 1: collision between two identical trains (single-unit train sets or defined formation) at a relative speed of 36 km/h.
- Scenario 2: collision between a train (single-units train sets or defined formation) and a railway vehicle equipped with side buffers at a speed of 36 km/h. The railway vehicle shall be a four-axle freight wagon with a mass of 80 t.
- Scenario 3: collision at a speed of 110 km/h, on a level crossing, with an obstacle equivalent to a 15 t specially defined lorry.
- Scenario 4: collision with a small or low obstacle such as a car or animal, which shall be addressed by defining the characteristics of an obstacle deflector.

To fulfill these scenarios a certain energy-absorbing capacity must be achieved by the couplers and the body shells. Figure 13.148 shows an example of energy absorption.

13.3.8 Air Conditioning

For high-speed rail vehicles closed body shells with fixed windows are essential for safety reasons. Air conditioning is therefore a must. For environmental reasons air is also used for cooling purposes (Figs. 13.149 and 13.150).

Process Air Loop. In the cooling circuit, the process air is first expanded in the motorized air cycle machine (MAM), and thus cooled. The cold process air passes through the downstream heat exchanger to chill the air supply to the passenger saloon. Subsequently, the now heated process air is taken in by the turbocompressor and released to the ambient.

The cooling process is controlled by the air cycle machine speed. The cooling power is thus infinitely variable between 0 and 100%.

Supply Air Loop. The supply air is a mixture of ambient fresh air and recirculated air from the vehicle. The supply air fan delivers the supply air through a supply air filter to the heat exchanger, where the required temperature reduction is performed in cooling mode.

The conditioned supply air is passed through an air silencer to the passenger compartment. In heating mode, the MAM is turned off so that the cooling process is deactivated. Controlled operation of the main heater adds the required amount of heat to the air flow.

To balance the air flow to the saloon, the equivalent amount of air to the fresh air quantity has to be exhausted. The cooling energy contained in the exhaust air will be used regeneratively in the process loop to improve the system's efficiency.

13.4 Aerospace Engineering

Aerospace engineering is a branch of engineering that deals with the design, construction, and operation of aerospace vehicles.

13.4.1 Aerospace Industry

The aerospace industry is a collection of organizations involved in research, design, construction, test, and operation of aerospace vehicles. In the USA, the aerospace industry consists of 20 prime contractors, 10 major airlines, a large government-supported research agency, and thousands of smaller companies that supply special components to these prime contractors. The total employment in the industry varies somewhat with changing business conditions, but in recent years has averaged about 2.5 million people, of whom approximately 75 000 are employed as engineers. The aerospace industries in developed countries such as Russia, Japan, and the European countries are not quite as large as that of the USA in terms of number of employees, but are similar in nearly every respect. Aerospace industries are growing rapidly in populous countries such as China and India.

Aerospace Industry Product Classifications

The products of the aerospace industry are many and varied, meeting a number of mission requirements. The broadest product classifications are related to the customer purchasing the product, giving rise to the classification into civil and military products [13.96]. More specific product classifications are derived from the type of aerospace vehicle and the particular use to which it is put. The section is organized in this manner.

13.4.2 Aircraft

The term *aircraft* is an all-inclusive term for any form of craft designed for navigation in the air. In the years since the first actual man-carrying flight in a hot-air balloon in the late 1700s, there have been a number of types of aircraft that have provided the means for aerial navigation. A brief recap includes the hot-air balloon ascension of de Rozier and d'Arlandes in 1783, the hydrogen balloon flight of J. A. C. Charles and M. N. Robert in 1783, and the successful steam-engine-powered airship (balloon) of Giffard in 1852. The German Otto Lilienthal developed a man-carrying glider that made over 2000 glides before suffering a fatal crash in 1896. As has been well documented, it was the Wright brothers, Orville and Wilbur, who made the first controlled powered flights of an airplane in 1903. Progress in aircraft design was slow in the first few years after the Wright flights, but by the start of World War I (WWI), many flying machines of various types and configurations had been successfully built and flown. During WWI, military requirements gave rise to the development of numerous types of aircraft with very specialized capabilities, which were produced in their thousands. Following the war, the use of aircraft for the transportation of passengers came into widespread use, with the establishment of airline companies and air routes, first between major European cities, but later in America and other parts of world as well. In the period between World War I and World War II, specialized aircraft were used to set nonstop distance records between continents, while other specialized aircraft set records for speed and altitude. World War II saw the introduction of new technology in aircraft design with the advent of practical helicopters, jet engines, rocket propulsion systems, and guided missiles. Following World War II, there was significant growth in private, recreational flying, expansion of the international commercial air transportation system, as well as continuing development of experimental aircraft that flew higher, faster, and farther than previous aircraft. With the creation of the National Aeronautics and Space Administration (NASA) in 1958, a variety of unique aircraft and spacecraft have been designed to meet very specific mission objectives laid down by that Agency. In recent years, there has been increasing military interest in unmanned combat air vehicles (UCAVs).

Aircraft Types

The two major categories of aircraft types are lighter than air (LTA) and heavier than air (HTA). A lighter than air craft is one that rises aloft by making use of Archimedes' principle, that is, by displacing a weight of air that is greater than the weight of the craft itself, and so creating a buoyant force. A heavier-than-air aircraft is one that rises aloft due to Bernoulli's principle acting on the aircraft's lifting surfaces, creating suctions on the upper surface and pressures on the lower surface relative to the ambient air pressure [13.97–127].

The design and operation of civil aircraft in the USA is subject to numerous regulations promulgated by the Federal Aviation Administration (FAA) of the Department of Transportation. Table 13.15 presents a summary of the various types of FAA regulations.

Table 13.15 Summary of Federal Aviation Regulatory (FAR) categories

Regulatory category	FAR part
Certification procedures for products and parts	21
Airworthiness standards, normal, utility, acrobatic, and commuters	23
Airworthiness standards, transport category airplanes	25
Airworthiness standards, normal category rotorcraft	27
Airworthiness standards, transport category rotorcraft	29
Airworthiness standards, manned free balloons	31
Airworthiness standards, aircraft engines	33
Airworthiness standards, propellers	35
Noise standards, aircraft type and airworthiness standards	36
Airworthiness directives	39
Maintenance, preventive maintenance, rebuilding, and alteration	43
Identification and registration marking	45
Aircraft registration	47
General operating and flight rules	91
Special air-traffic rules and airport traffic patterns	93
IFR (instrument flight rules) altitudes	95
Standard instrument approach procedures	97
Ultralight vehicles	103
Certification and operation, domestic, flag, and supplemental air carriers, and commercial operators of large aircraft	121
Certification and operation, aiplanes having seating capacity of 20 or more passengers, or a maximum payload capacity of 6,000 pounds or more	125
Certification and operation of scheduled air carrier helicopters	127
Air taxi operators and commercial operators	135
Agricultural aircraft operations	139

Lighter-Than-Air Aircraft. One can distinguish between the following types of LTA aircraft:

- Hot-air balloon, which consist of a large envelope made of lightweight fabric to contain the hot air, a burner located below the envelope, usually fueled by kerosene, to heat ambient air, causing it to rise into the envelope. A basket hung underneath the burner is provided for the pilot and passengers.
- Light-gas balloon, which is similar in arrangement to a hot-air balloon, but without the burner. The buoyant force is generated by the use of light gasses such as helium in the envelope, which displace the relatively heavier ambient air.
- Blimp or nonrigid airship, which is basically a large gas balloon whose streamlined shape is maintained by internal gas pressure. In addition to the gas envelope, the blimp has a car attached to the lower part of the envelope for the crew and passengers, engines and propellers to develop forward speed, and fins with hinged aft portions for control.
- Rigid airship, a lighter-than-air aircraft with a rigid frame to maintain its shape and provide a volume for the internal placement of light gasbags. The rigid airship also has a car attached to the lower part of the rigid frame for the crew and passengers, engines, propellers, and tail fins similar to the blimp. Rigid airships reached the peak of their development in the mid 1930s, but several spectacular accidents curtailed further development.

Heavier-Than-Air Craft. HTA aircraft can be divided into three main categories.

A *glider* is an aircraft that flies without an engine. The simplest form of a glider is the hang glider, which consists of a wing, a control frame, and a pilot harness. The pilot is zipped into the harness and literally hangs beneath the wing, with his hands on the control bar of the control frame. The wing has an aluminum frame that supports the wing fabric, and internal battens to provide a proper shape to the fabric. Hang gliders are usually launched from a very steep hill or a cliff that affords sufficient altitude for gliding flight. Simple utility gliders which have rigid structural elements similar to an airplane are used primarily for training. These gliders are launched into the air by being towed by a power winch, an automobile, or an airplane. Extremely refined sailplanes, usually made of very lightweight materials and featuring very long thin wings, take advantage of rising air currents, and can soar for a long time and cover distances of hundred of miles in a single flight. A variation of the sailplane is the *motorglider*, basically a sailplane with a small motor and propeller, which is used for take-off and climb to soaring altitude, whereupon the motor is shut off, and is then retracted along with the propeller to revert to the sailplane configuration.

An *airplane* is an air vehicle that incorporates a propulsion system and fixed wings, and is supported by aerodynamic forces acting on the wings. Airplane propulsion systems may be a piston engine driving a propeller, a turbojet engine, or a rocket engine, depending on the required mission. Airplanes range in form from small general aviation aircraft, usually privately owned, with one or two engines, to larger commercially operated air transport aircraft that can carry from 20 to upwards of 500 passengers and can fly distances from 500 to over 8000 miles nonstop. In addition to these civil aircraft types, there a number of military aircraft types designed for different missions, such as fighter, attack, bomber, reconnaissance, transport, and trainer. A very small class of airplanes, known as experimental research aircraft, usually powered by rocket engines, has been built to obtain flight test data at extremely high speeds and altitudes. The ultimate development in this area is the US Space Shuttle, which is a rocket-powered spacecraft for most of its mission, and an unpowered glider for the approach and landing phase of the flight.

A *helicopter* is an aircraft that is supported by aerodynamic forces generated by long thin blades rotating about a vertical axis. The rotor blades are driven by the helicopter's propulsion system, usually a piston or gas turbine engine. Helicopters range in size from small, two-seat personal utility models to large transport types that can carry up to 40 people. Large heavy-lift helicopters are often used in specialized hauling and construction tasks, where their ability to remain airborne over a fixed spot for extended periods of time is unique. Helicopters have also been used in several military applications such as air–sea rescue, medical evacuation, as battlefield gunships, and for special-operations troop transport. Recent developments in helicopter technology have led to hybrid helicopter craft called the tilt rotor. In this machine, there are two rotors to provide the vertical forces required for take-off and landing, but as the name implies, these rotors may be tilted to varying degrees until they are aligned in the direction of flight, acting like the propellers on a conventional airplane. The tilt rotor has small wings to provide the aerodynamic lift required during cruise flight, during which the rotors are used to provide forward thrust.

13.4.3 Spacecraft

Spacecraft fall into two major categories, unmanned, with no humans aboard, and manned, with humans aboard. Examples of unmanned spacecraft include civil communication satellites, military reconnaissance satellites, and scientific probes that gather information on our solar system. Examples of unmanned spacecraft include the Echostar and Eutelsat civil communication satellites, the Aquila and Cosmos military reconnaissance satellites, and the Hubble Space Telescope and Mars Global Surveyor scientific probes. Examples of manned spacecraft include the Vostok, Soyuz, Mir, Mercury, Gemini, Apollo, and Space Shuttle vehicles.

13.4.4 Definitions

The following are some important definitions related to a good understanding of aerospace engineering.

Units

Although there has been a policy in the USA in recent years to convert to the international system (SI) of units, the US aerospace industry continues to use English units in its work. This publication will use English units as primary, since most American engineers are familiar with this terminology. A list of conversion factors between SI and English units is given in Table 13.16.

Table 13.16 Conversion factors between SI and English units

Conversion factors	
Mass	$1.00\,\text{kg} = 0.06853\,\text{slug}$ $1.00\,\text{slug} = 14.592\,\text{kg}$ At the surface of the Earth, an object with a mass of 1.00 kg weighs 9.8 N or 2.205 lb, and an object with a mass of 1.00 slug weighs 32.17 lb or 143.1 N
Length	$1.00\,\text{m} = 3.2808\,\text{ft}$ $1.00\,\text{ft} = 0.3048\,\text{m} = 30.48\,\text{cm}$
Force	$1.00\,\text{N} = 0.2248\,\text{lb}$ $1.00\,\text{lb} = 4.4482\,\text{N}$
Temperature	$1.00\,\text{K} = 1.8\,^\circ\text{Ra}$ $\quad ^\circ\text{Ra} = ^\circ\text{F} + 460$ $1.0\,^\circ\text{Ra} = 0.5556\,\text{K}$ $\quad \text{K} = ^\circ\text{C} + 273$
Pressure	$1.00\,\text{N/m}^2 = 1.4504 \times 10^{-4}\,\text{lb/in}^2 = 2.0886 \times 10^{-2}\,\text{lb/ft}^2$ $1.00\,\text{lb/in}^2 = 6.8947 \times 10^3\,\text{N/m}^2$ $1.00\,\text{lb/ft}^2 = 47.88\,\text{N/m}^2$
Velocity	$1.00\,\text{m/s} = 3.2808\,\text{ft/s} = 2.2369\,\text{mi/h}$ $1.00\,\text{ft/s} = 0.6818\,\text{mi/h} = 0.3048\,\text{m/s}$
Density	$1.00\,\text{kg/m}^3 = 1.9404 \times 10^{-3}\,\text{slug/ft}^3$ $1.00\,\text{slug/ft}^3 = 515.36\,\text{kg m}^3$
Viscosity	$1.00\,\frac{\text{kg}}{\text{m s}} = 20.886 \times 10^2\,\frac{\text{lb s}}{\text{ft}^2}$ $1.00\,\frac{\text{lb s}}{\text{ft}^2} = 47.879\,\frac{\text{kg}}{\text{m s}}$
Specific heat	$1.00\,\frac{\text{N m}}{\text{kg K}} = 1.00\,\frac{\text{J}}{\text{kg K}} = 2.3928\,\frac{\text{BTU}}{\text{lb}_\text{m}\,^\circ\text{Ra}} = 5.9895\,\frac{\text{ft lb}_\text{f}}{\text{slug}\,^\circ\text{Ra}}$ $1.00\,\frac{\text{ft lb}}{\text{slug}\,^\circ\text{Ra}} = 1.6728 \times 10^{-1}\,\frac{\text{N m}}{\text{kg K}} = 1.6728 \times 10^{-1}\,\frac{\text{J}}{\text{kg K}}$
Frequently used equivalents	
1 bhp	550 ft lb/s = 33 000 ft lb/min
1 knot (kn) (i. e., nautical mile per hour)	1.152 statute mile per hour
1 knot (kn) (nautical mile per hour)	1.69 ft/s
1 statute mile per hour	0.868 knot (nautical miles per hour)
1 statute mile per hour	1.467 ft/s
1 ft/s	0.682 statute mile per hour
1 ft/s	0.592 knot (nautical miles per hour)
1 kilometer	0.621 statute mile
1 kilometer	0.539 nautical mile
1 statute mile	1.609 kilometer
1 nautical mile	1.854 kilometer
1 radian	57.3 degrees
Note that the preceeding values are *equivalents*. The conversion factors are the reciprocals.	
Frequently used constants	
γ	1.4 (air)
Gas constant R (air)	287.05 N m/(kg K) = 1718 ft lb/(slug °Ra)
Specific heat c_p (air)	1004.7 N m/(kg K) (J/(kg K)) = 6006 ft lb/(slug °Ra)
Gravitational constant at sea level g_0	$9.8\,\text{m/s}^2 = 32.17\,\text{ft/s}^2$
Radius of the Earth r_0	$6.378 \times 10^6\,\text{m} = 20.92 \times 10^6\,\text{ft}$

Flight Speed Terminology

One of the key performance parameters for an airplane is its maximum level-flight speed. For a variety of technical and economic reasons, various airplanes are designed to operate at speeds most appropriate to their design missions. Modern airplanes operate at speeds

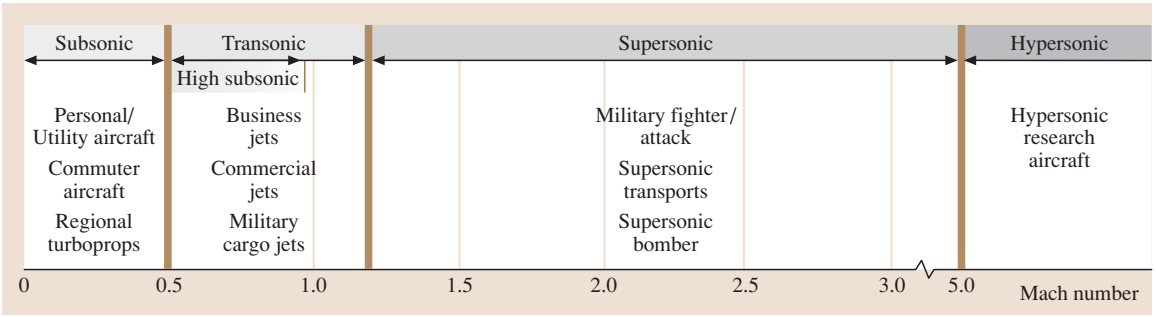


Fig. 13.151 Flight speed terminology

ranging from a low of around 60 kn to highs of around 1450 kn. Over such a wide range of flight speeds, the characteristics of the airflow around the airplane change dramatically. These changes, associated with the compressible nature of air, are directly related to the flight Mach number, defined as the flight speed divided by the speed of sound in the ambient air in which the airplane is flying. This situation has given rise to some general terms to describe airplane flight speeds in terms of Mach number, as shown in Fig. 13.151. Also shown are the types of airplanes having maximum level-flight speeds within the various speed regimes.

Standard Atmosphere

For design and performance calculations, it is appropriate to establish a standard set of characteristics for the Earth's atmosphere in which aircraft operate. The US standard atmosphere is a widely used set whose essential characteristics, that is, the temperature, pressure, density, and viscosity, as a function of altitude have been derived using

$$p = \rho RT\,,$$
$$\mathrm{d}p = -\rho g\,\mathrm{d}h\,,$$

where

p = pressure in lb/ft^2,
ρ = density in slug/ft,
T = absolute temperature in °Ra,
R = gas constant (1718 ft lb/(slug °Ra)) for air,
g = gravitational constant (32.17 ft/s^2),
H = height above sea level in ft.

With these equations, only a defined variation of T with altitude is required to establish the standard atmosphere. The defined variation, based on experimental data, is shown in Fig. 13.152.

Once the temperature variation with altitude is defined, the characteristics of the standard atmosphere can be calculated directly.

The characteristics of the US standard atmosphere are tabulated in Table 13.17. From sea level to 36 089 ft the temperature decreases linearly with altitude. This region is called the troposphere. Above 36 089 ft the temperature is constant up to 65 617 ft in the region called the stratosphere. Above 65 617 ft the temperature increases linearly to 100 000 ft, the upper level of interest for current or foreseeable aircraft.

Although the concept of geometric altitude, the altitude above sea level as determined by a tape measure, is most familiar, of prime importance for aircraft design and performance calculations is the pressure altitude, i.e., the geometric altitude on a standard day for which the pressure is equal to the ambient atmospheric pres-

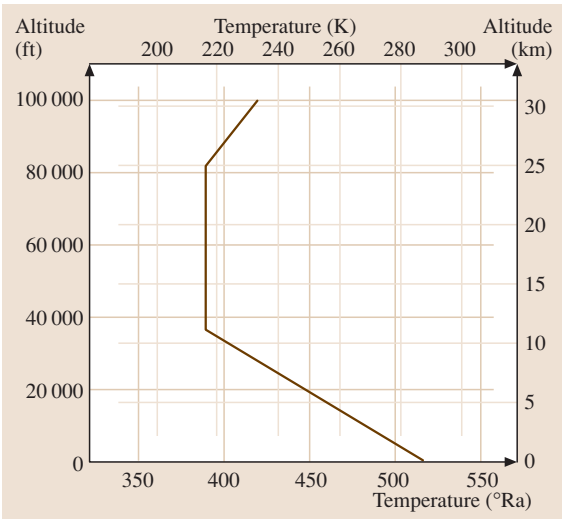


Fig. 13.152 Temperature variation with altitude in the US standard atmosphere

Table 13.17 Characteristics of the US standard atmosphere

Altitude	Temperature		Pressure	Density	Density ratio	Kinematic viscosity	q/M^2	Sonic velocity	
(ft)	(°F)	(°Ra)	(psf)	(slug/cu ft)	(σ)	(ft^2/s)	(lb/ft^2)	(ft/s)	(kn)
0	59.0	518.7	2116.2	0.0023769	1.0000	0.0001572	1481.0	1116.4	661.5
1000	55.4	515.1	2040.9	0.0023081	0.9710	0.0001610	1429.0	1112.6	659.2
2000	51.9	511.6	1967.7	0.0022409	0.9427	0.0001650	1377.0	1108.7	656.9
3000	48.3	508.0	1896.7	0.0021752	0.9151	0.0001691	1328.0	1104.9	654.6
4000	44.7	504.4	1827.7	0.0021110	0.8881	0.0001732	1279.0	1101.0	652.3
5000	41.2	500.9	1760.9	0.0020482	0.8616	0.0001776	1233.0	1097.1	650.0
6000	37.6	497.3	1696.0	0.0019869	0.8358	0.0001820	1187.0	1093.2	647.7
7000	34.0	493.7	1633.1	0.0019270	0.8106	0.0001866	1143.0	1089.3	645.4
8000	30.5	490.2	1572.1	0.0018685	0.7860	0.0001914	1100.0	1085.3	643.0
9000	26.9	486.6	1512.9	0.0018113	0.7619	0.0001963	1059.0	1081.4	640.7
10 000	23.3	483.0	1455.6	0.0017556	0.7385	0.0002013	1019.0	1077.4	638.3
11 000	19.8	479.5	1400.0	0.0017011	0.7155	0.0002066	979.8	1073.4	636.0
12 000	16.2	475.9	1346.2	0.0016480	0.6932	0.0002120	942.1	1069.4	633.4
13 000	12.6	472.4	1294.1	0.0015961	0.6713	0.0002175	905.6	1065.4	631.4
14 000	9.1	468.8	1243.6	0.0015455	0.6500	0.0002233	870.2	1061.4	628.8
15 000	5.5	465.2	1194.8	0.0014962	0.6292	0.0002293	836.0	1057.4	626.4
16 000	1.9	461.7	1147.5	0.0014480	0.6089	0.0002354	802.9	1053.3	624.0
17 000	−1.6	458.1	1101.7	0.0014011	0.5892	0.0002418	770.8	1049.2	621.6
18 000	−5.2	454.6	1057.5	0.0013553	0.5699	0.0002484	739.8	1045.1	619.2
19 000	−8.8	451.0	1014.7	0.0013107	0.5511	0.0002553	709.8	1041.0	616.7
20 000	−12.3	447.4	973.3	0.0012673	0.5328	0.0002623	680.8	1036.9	614.3
21 000	−15.9	443.9	933.3	0.0012249	0.5150	0.0002697	652.7	1032.8	611.9
22 000	−19.5	440.3	894.6	0.0011836	0.4976	0.0002772	625.6	1028.6	609.4
23 000	−23.0	436.8	857.2	0.0011435	0.4806	0.0002851	599.4	1024.5	606.9
24 000	−26.6	433.2	821.2	0.0011043	0.4642	0.0002932	574.1	1020.3	604.4
25 000	−30.2	429.6	786.3	0.0010663	0.4481	0.0003017	549.7	1016.1	601.9
26 000	−33.7	426.1	752.7	0.0010292	0.4325	0.0003104	526.2	1011.9	599.4
27 000	−37.3	422.5	720.3	0.0009931	0.4173	0.0003195	503.4	1007.7	596.9
28 000	−40.9	419.0	689.0	0.0009580	0.4025	0.0003289	481.5	1003.4	594.4
29 000	−44.3	415.4	658.8	0.0009239	0.3881	0.0003387	460.3	999.1	591.9
30 000	−48.0	411.9	629.7	0.0008907	0.3741	0.0003488	439.9	994.8	589.3
31 000	−51.6	408.3	601.6	0.0008584	0.3605	0.0003594	420.3	990.5	586.8
32 000	−55.1	404.8	574.6	0.0008270	0.3473	0.0003703	401.3	986.2	584.2
33 000	−58.7	401.2	548.5	0.0007966	0.3345	0.0003817	383.1	981.9	581.6
34 000	−62.3	397.6	523.5	0.0007670	0.3220	0.0003935	365.5	977.5	579.0
35 000	−65.8	394.1	499.3	0.0007382	0.3099	0.0004058	348.6	973.1	576.4
36 000	−69.4	390.5	476.1	0.0007103	0.2981	0.0004185	332.3	968.8	573.8
37 000	−69.7	390.0	453.9	0.0006780	0.2843	0.0004379	330.9	968.1	573.6
38 000	−69.7	390.0	432.6	0.0006463	0.2710	0.0004594	316.7	968.1	573.6
39 000	−69.7	390.0	412.4	0.0006161	0.2583	0.0004820	301.8	968.1	573.6
40 000	−69.7	390.0	393.1	0.0005873	0.2462	0.0005056	287.7	968.1	573.6
41 000	−69.7	390.0	374.6	0.0005598	0.2346	0.0005304	274.2	968.1	573.6

Table 13.17 (cont.)

Altitude	Temperature		Pressure	Density	Density ratio	Kinematic viscosity	q/M^2	Sonic velocity	
(ft)	(°F)	(°Ra)	(psf)	(slug/cu ft)	(σ)	(ft^2/s)	(lb/ft^2)	(ft/s)	(kn)
42 000	−69.7	390.0	357.2	0.0005336	0.2236	0.0005564	261.3	968.1	573.6
43 000	−69.7	390.0	340.5	0.0005087	0.2131	0.0005837	249.0	968.1	573.6
44 000	−69.7	390.0	324.6	0.0004849	0.2031	0.0006123	237.4	968.1	573.6
45 000	−69.7	390.0	309.4	0.0004623	0.1936	0.0006423	226.2	968.1	573.6
46 000	−69.7	390.0	295.0	0.0004407	0.1845	0.0006738	215.6	968.1	573.6
47 000	−69.7	390.0	281.2	0.0004201	0.1758	0.0007068	205.5	968.1	573.6
48 000	−69.7	390.0	268.1	0.0004004	0.1676	0.0007415	195.8	968.1	573.6
49 000	−69.7	390.0	255.5	0.0003818	0.1597	0.0007778	186.7	968.1	573.6
50 000	−69.7	390.0	243.6	0.0003639	0.1522	0.0008159	177.9	968.1	573.6
51 000	−69.7	390.0	232.2	0.0003469	0.1451	0.0008559	169.5	968.1	573.6
52 000	−69.7	390.0	221.4	0.0003307	0.1383	0.0008978	161.6	968.1	573.6
53 000	−69.7	390.0	211.0	0.0003153	0.1318	0.0009418	154.0	968.1	573.6
54 000	−69.7	390.0	201.2	0.0003006	0.1256	0.0009879	146.8	968.1	573.6
55 000	−69.7	390.0	191.8	0.0002865	0.1197	0.0010360	139.9	968.1	573.6
56 000	−69.7	390.0	182.8	0.0002731	0.1141	0.0010871	133.3	968.1	573.6
57 000	−69.7	390.0	174.3	0.0002604	0.1087	0.0011403	127.1	968.1	573.6
58 000	−69.7	390.0	166.2	0.0002482	0.1036	0.0011961	121.1	968.1	573.6
59 000	−69.7	390.0	158.4	0.0002366	0.0988	0.0012547	115.4	968.1	573.6
60 000	−69.7	390.0	151.0	0.0002256	0.0841	0.0013161	110.0	968.1	573.6
61 000	−69.7	390.0	144.0	0.0002151	0.0897	0.0013805	104.8	968.1	573.6
62 000	−69.7	390.0	137.3	0.0002050	0.0855	0.0014481	99.9	968.1	573.6
63 000	−69.7	390.0	130.9	0.0001955	0.0815	0.0015189	95.2	968.1	573.6
64 000	−69.7	390.0	124.8	0.0001834	0.0777	0.0015932	90.8	968.1	573.6
65 000	−69.7	390.0	118.9	0.0001777	0.0740	0.0016712	86.5	968.1	573.6
70 000	−67.3	392.4	92.7	0.0001376	0.0579	0.0021219	82.4	971.0	575.3
75 000	−64.6	395.1	73.0	0.0001077	0.0453	0.0026938	64.9	974.4	577.3

sure. Aircraft altimeters are pressure gages calibrated to read pressure altitude. Also important is the density altitude, the geometric altitude on a standard day for which the density is equal to the ambient air density. Pressure altitude, density altitude, and temperature are related through the equation of state $p = \rho RT$.

It should be noted that another standard atmosphere has been defined by the International Civil Aviation Organization (ICAO). The ICAO standard atmosphere and the US standard atmosphere are identical up to 65 617 ft. Beyond 65 617 ft the ICAO standard atmosphere maintains a constant temperature up to 82 300 ft, while the US standard atmosphere reflects an increasing temperature with a constant gradient to beyond 100 000 ft.

Axis Systems

The airplane design process, with respect to achieving performance objectives of altitude, speed, range, payload, and take-off and landing distance requires analysis of the airplane in motion. The Newtonian laws of motion state that the summation of all external forces in any direction must equal the time rate of change of momentum, and that the summation of all of the moments of the external forces must equal the time rate of change of moment of momentum, all measured with respect to axes fixed in space. If the motion of the airplane is described relative to axes fixed in space, the mathematics becomes extremely unwieldy, as the moments and products of inertia vary from instant to instant. To overcome this difficulty, use is made of moving or Eulerian axes that coincide in some particular manner from instant to instant with a definite set of axes fixed with respect to the airplane. The most common choice is to select a set of mutually perpendicular axes defined within the airplane as shown in Fig. 13.153, with their origin at the airplane center of gravity (c.g.). The airplane's motion

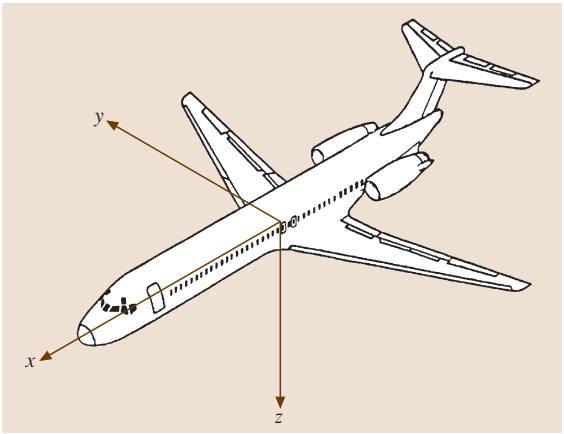


Fig. 13.153 Airplane axis system

in space is defined by six components of velocity. This is a right-hand axis system with the positive x- and z-axes in the plane of symmetry and with the x-axis out the nose of the airplane pointing along the flight path. This is called a wind axis system, since the x-axis always coincides with the airplane's velocity vector. The z-axis is perpendicular to the x-axis, positive downward, and the positive y-axis is out the right hand wing, perpendicular to the plane of symmetry.

For most aircraft design analyses, the airplane is considered to be a rigid body with six degrees of freedom: three linear velocity components along these axes, and three angular velocity components around these axes. The angular motion around the y-axis is called pitch; the angular motion about the x-axis is called roll; and the angular motion about the z-axis is called yaw. Nearly all of the airplane motions encountered in aircraft preliminary design and performance are in the plane of symmetry. The other three components of the airplane's motion lie outside the plane of symmetry. The symmetric degrees of freedom are referred to as the longitudinal motion, and the asymmetric degrees of freedom are referred to as the lateral-directional motion.

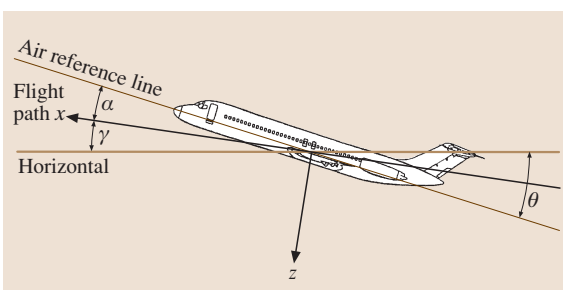


Fig. 13.154 Axis system in the plane of symmetry

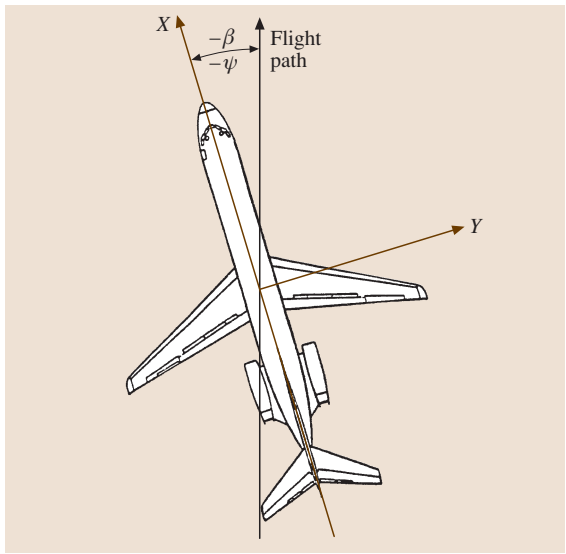


Fig. 13.155 Axis system in asymmetric flight

In the plane of symmetry (Fig. 13.154) the inclination of the flight path to the horizontal is the flight path angle γ and the angle between the flight path and the airplane reference line is the angle of attack α. The angle between the airplane reference line and the horizontal is the airplane's pitch angle θ.

When the flight path does not lie in the plane of symmetry (Fig. 13.155) the angle between the flight path and the airplane's centerline is the yaw angle ψ. For straight flight in this situation, the yaw angle is equal in magnitude but opposite in sign to the sideslip angle β. In roll about the flight path, the angle between the y-axis and the horizontal is the roll or bank angle. A summary of all of the angles involved in flight performance and flight mechanics calculations is presented in Table 13.18.

The other three components of the airplane's motion lie outside the plane of symmetry. The symmetric degrees of freedom are referred to as the longitudinal motion, and the asymmetric degrees of freedom are referred to as the lateral-directional motion. In the plane of symmetry (Fig. 13.154) the inclination of the flight path to the horizontal is the flight path angle γ and the angle between the flight path and the airplane reference line is the angle of attack α.

The angle between the airplane reference line and the horizontal is the airplane's pitch angle θ. When the flight path does not lie in the plane of symmetry

Table 13.18 Summary and definition of the angles involved in flight performance and flight mechanics calculations

In the plane of symmetry		
γ	Flight path angle	Angle between the horizon and the velocity vector
θ	Pitch angle	Angle between the airplane reference line and the horizon
α	Angle of attack	Angle between the airplane reference line and the velocity vector
In asymmetric flight		
β	Angle of sideslip	Angle between the airplane centerline and the velocity vector
$\varnothing$	Angle of roll	Angle between the airplane's y-axis and the horizon

(Fig. 13.155) the angle between the flight path and the airplane's centerline is the yaw angle ψ. For straight flight in this situation, the yaw angle is equal in magnitude but opposite in sign to the sideslip angle β. In roll about the flight path, the angle between the y-axis and the horizontal is the roll or bank angle. A summary of all of the angles involved in flight performance and flight mechanics calculations are presented in Table 13.18.

Aerodynamic Forces and Moments

The aerodynamic forces acting on an aircraft consist of two types: pressure forces, which act normal to the aircraft surface, and viscous or shear forces, which act tangentially to the aircraft surface.

The physical parameters that govern the aerodynamic forces and moments acting on an aircraft have been developed through a method called dimensional analysis. This procedure is treated in detail in many textbooks on aerodynamics, and will only be summarized here. Dimensional analysis considers the dimensions or units of the physical quantities involved in the development of aerodynamic forces and moments, and divides them into two groups: fundamental and derived. The fundamental units are mass, length, and time, and all physical quantities have dimensions that are derived from a combination of these three fundamental units. Equations that express physical relationships must have dimensional homogeneity; that is, each term in the equation must have the same units in order for the equation to have physical significance. The broad physical relationships are postulated by logic, reason, or perhaps some experimental evidence, and then the specific relationships are derived by dimensional analysis. For aerodynamic forces and moments acting on an aircraft in the plane of symmetry the broad physical relationships are postulated as

$$F_{\text{aero}},\ M_{\text{aero}} = f(\text{shape, size, altitude, velocity, fluid properties}) .$$

The specific relationship for aerodynamic forces, derived from dimensional analysis, is

$$F_{\text{aero}} = K\rho V^2 L^2 f\left(\alpha, \frac{\rho VL}{\mu}, \frac{V}{a}\right) ,$$

where

K	is a constant of proportionality or dimensionless coefficient,
V	is the velocity of the aircraft,
L	is an arbitrary characteristic length,
ρ	is the air density,
μ	is the coefficient of viscosity for air,
α	is the altitude of the airplane with respect to the flight path,
$\rho VL/\mu$	is a dimensionless quantity called the Reynolds number Re,
a	is the speed of sound in air,
V/a	is a dimensionless quantity called the Mach number Ma.

The aerodynamic forces and moments acting on the airplane in the plane of symmetry are shown in Fig. 13.156. The resultant of the aerodynamic forces is resolved into the lift component, acting perpendicular to the flight path or velocity vector, and the drag component, acting parallel to velocity vector. The lift and

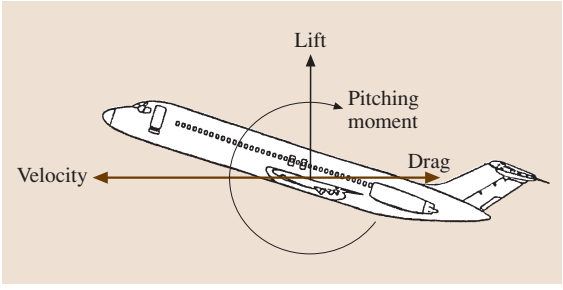


Fig. 13.156 Aerodynamic forces and moments in the plane of symmetry

drag components are defined as acting at the airplane center of gravity, while all of the moments acting on the airplane are lumped into one couple acting around the airplane center of gravity. The equations for lift and drag of the airplane may be written as

$$\text{lift} = C_L \frac{\rho}{2} V^2 S \,,$$

$$\text{drag} = C_D \frac{\rho}{2} V^2 S \,,$$

where C_L and C_D are the lift and drag coefficients, respectively, and the area term in the equation is arbitrarily taken to be the wing area, S. To make the equation for the aerodynamic moment about the center of gravity dimensionally correct, the length of the wing mean chord, c, the mean distance from the leading edge to the trailing edge at the wing is arbitrarily selected. The moment equation in the plane of symmetry is

$$\text{Moment} = M_{cg} = C_m \frac{\rho}{2} V^2 S c \,,$$

where C_m is defined as the pitching moment coefficient. As noted, while the primary relationship between the physical quantities involved in the development of aerodynamic forces and moments is expressed in terms of the dimensionless coefficients, these coefficients are functions of both the Reynolds and the Mach number. The aerodynamic forces and moments acting on the airplane in asymmetric flight are shown in Fig. 13.157. The side force acts normal to the airplane center line, while the aerodynamic moments acting around the z-axis through the c.g. are lumped together and called the yawing moment. In addition, the aerodynamic moments acting around the x-axis are lumped together and are called the rolling moment.

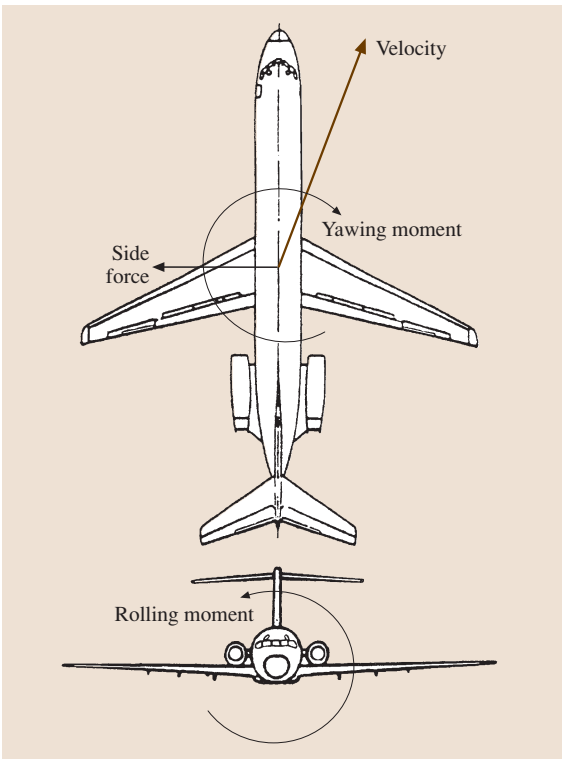


Fig. 13.157 Aerodynamic forces and moments in asymmetric flight

The equations for side force, yawing moment, and rolling moment are

$$\text{side force} = C_\gamma \frac{\rho}{2} V^2 S \,,$$

$$\text{yawing moment} = C_n \frac{\rho}{2} V^2 S b \,,$$

$$\text{side force} = C_l \frac{\rho}{2} V^2 S b \,,$$

where C_γ, C_n, and C_l are the side force, yawing moment, and rolling moment coefficients, respectively, and b is the airplane wing span, selected as more appropriate than the wing mean chord for use with the asymmetric moment coefficients.

In summary, then, there are three defined aerodynamic forces acting along the airplane axes, and three aerodynamic moments acting around the airplane axes (Table 13.19).

Relative Wind

Up to now, the aerodynamic forces acting on the airplane have been defined in terms of the airplane velocity vector. It should be noted that the aerodynamic forces and moments depend only on the relative velocity between the airplane and the air that it is flying through. The same aerodynamic forces are generated if the airplane moves through the air with a velocity, V, or if the airplane is held fixed in space, as in a wind tunnel, and the air moves past the airplane, coming from the direction of the relative wind, opposite to the flight path, with a velocity, V, equal and opposite to the actual velocity, as shown in Fig. 13.158.

Table 13.19 Aerodynamic forces and moments acting along the airplane axis

Axis	Force along	Moment around
x	Drag	Rolling moment
y	Side force	Pitching moment
z	Lift	Yawing moment

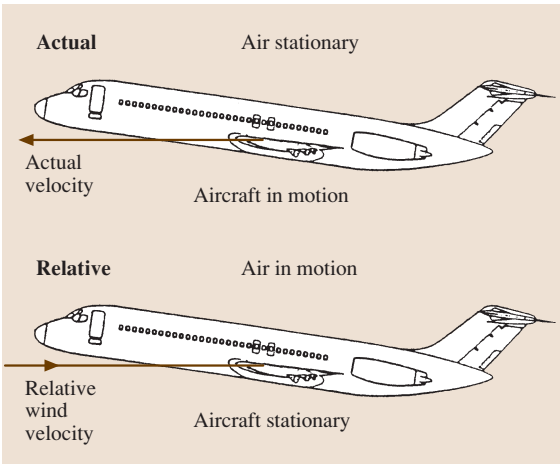


Fig. 13.158 Relative velocity and relative wind

Dynamic Pressure

In the discussion of aerodynamic forces and moments, the expressions for all of them show a dependency on the quantity $(\rho V^2)/2$. This quantity, which appears throughout aerodynamic theory, is equal to the kinetic energy of a unit volume of air, and is defined as the dynamic pressure:

$$q = \frac{\rho}{2} V^2 \, .$$

Another form of the equation for dynamic pressure which is especially useful in aircraft design and performance calculations is

$$q = \frac{\gamma}{2} p \mathrm{Ma}^2 \, ,$$

where

γ is the ratio of specific heats for air, equal to 1.4,
p is the ambient pressure, and
Mais the flight Mach number.

Airspeed Terminology

Since the very early days, airplanes have been equipped with airspeed indicators, which are operated by the pressure difference between two pressures sensed by devices mounted on the airplane. One pressure used in airspeed measurement is the impact or total pressure, usually sensed by a Pitot or total head tube (Fig. 13.159a), which has an open hole at the front end to capture the total pressure. The other pressure used is the static pressure, that is, the ambient pressure at the operating altitude of the airplane, usually sensed by small flush holes located on a static tube (Fig. 13.159b). On many airplanes, the Pitot tube and the static tube are integrated into one device called the Pitot-static tube (Fig. 13.159c). On larger aircraft, the static pressure is sensed by flush holes in the fuselage called static ports

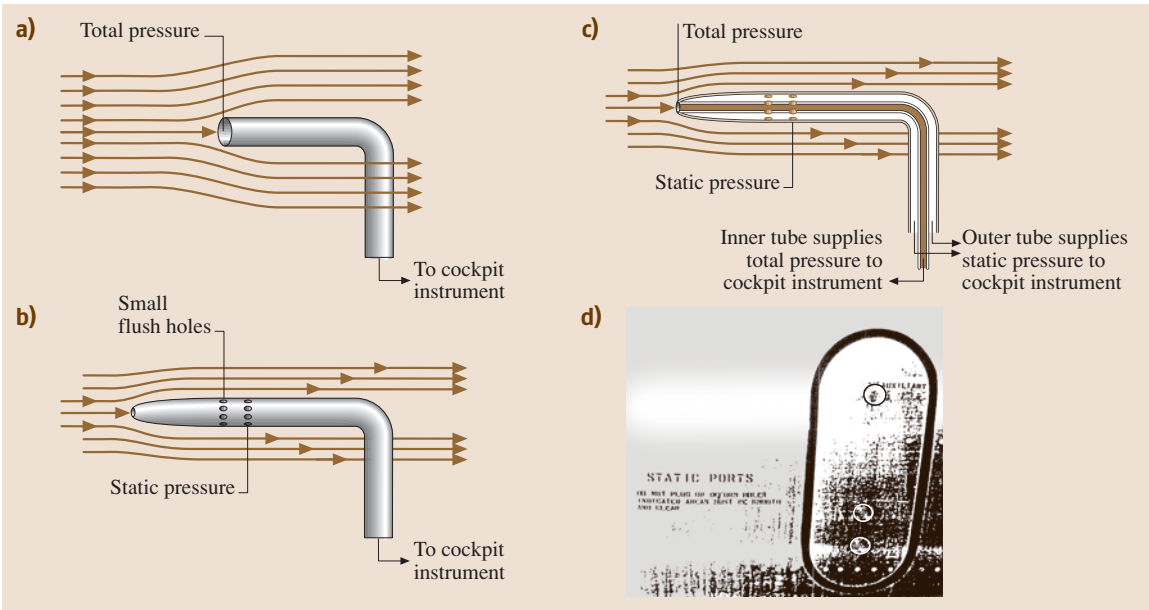


Fig. 13.159a–d Airspeed sensors

(Fig. 13.159d), which are located in an area where the static pressure is equal to the ambient static pressure. Airspeed indicators are calibrated to read airspeed as a function of the difference between total and static pressure. In order to develop some meaningful definitions of airspeed over a range of aircraft operational altitudes, some specific terminology has been adopted as follows.

Indicated Airspeed (IAS). This is the airspeed registered on the cockpit instrument, corrected for any instrument calibration errors.

Calibrated Airspeed (CAS). This is the indicated airspeed reading corrected for static source position errors. Usually it is not possible to locate a Pitot–static tube or a static port in a place that senses the exact ambient static pressure, so corrections for this so-called position error are made. Calibrated airspeed refers to the fact that airspeed indicators are calibrated in airspeed units, usually knots, through the difference between total and static pressure at sea-level standard day conditions by the equation

$$V_{\text{cal}} = \sqrt{\frac{2(P_{\text{total}} - P_{\text{static}})}{\rho_{\text{sea-level}}}} V_{\text{true}} \,.$$

Equivalent Airspeed (EAS). This is the calibrated airspeed adjusted for what are termed compressibility effects. Equivalent airspeed is a very important parameter in aircraft preliminary design and performance calculations. Equivalent airspeed is defined as

$$V_{\text{EAS}} = \sqrt{\frac{\rho_{\text{ambient}}}{\rho_{\text{sea-level}}}} V_{\text{true}} \,.$$

The idea behind equivalent airspeed is that, for any flight condition, i. e., any combination of true airspeed and ambient density, and therefore dynamic pressure, there is an equivalent airspeed at sea-level standard day conditions that produces the same dynamic pressure. In equation form

$$q = \frac{\rho_{\text{ambient}}}{2} V_{\text{true}}^2 = \frac{\rho_{\text{sea-level}}}{2} V_{\text{EAS}}^2 \,.$$

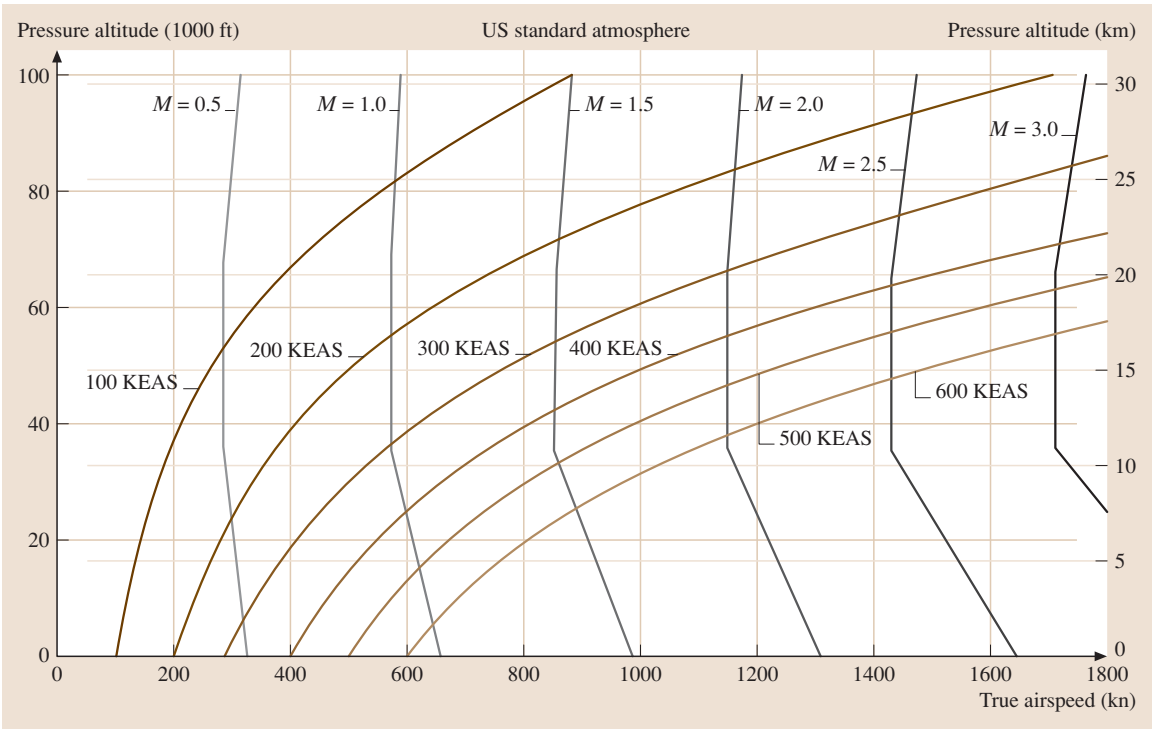


Fig. 13.160 V_{EAS} and V_{true}

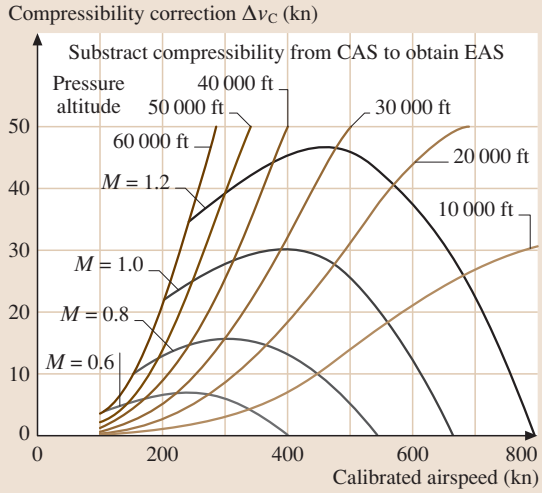


Fig. 13.161 Airspeed compressibility correction

A chart showing the relationship between V_{EAS} and V_{true} for the US standard atmosphere is shown in Fig. 13.160. Equivalent airspeeds are reasonably close to the indicated airspeeds shown by the cockpit instrument.

The difference is the compressibility correction to CAS in order to obtain EAS. The compressibility increment, ΔV_C, is a function of both the Mach number and altitude, as shown in Fig. 13.161. This correction arises as follows.

For a compressible fluid such as air,

$$(P_{total} - P_{static}) = \frac{\gamma}{2} p \mathrm{Ma}^2 \times \left(1 + \frac{\mathrm{Ma}^2}{4} + \frac{\mathrm{Ma}^4}{40} + \frac{\mathrm{Ma}^6}{1600} + \ldots\right) .$$

Since CAS is directly related to $(P_{total} - P_{static})$, and EAS is directly related to the dynamic pressure q, the correction is the Mach number series in the brackets.

True Airspeed (TAS). The true airspeed is the actual airspeed of the airplane at ambient conditions in the atmosphere, and may be obtained by converting or correcting equivalent airspeed as follows

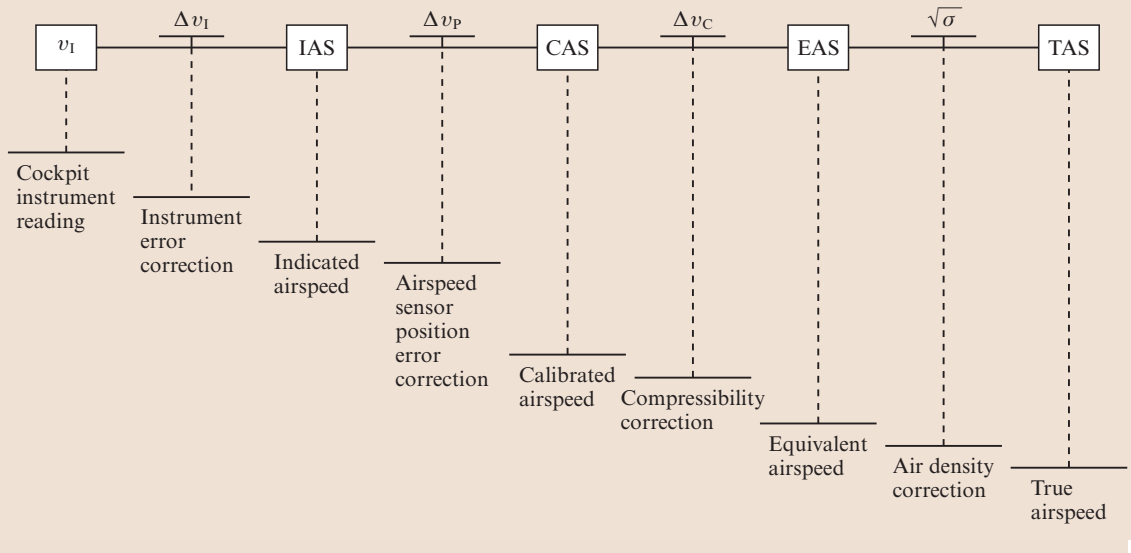
$$V_{true} = V_{EAS} \sqrt{\frac{\rho_{sea\text{-}level}}{\rho_{ambient}}} .$$

A summary of the airspeed indicator correction sequence is presented in Table 13.20.

13.4.5 Flight Performance Equations

In order to examine the major characteristics of the airplane's performance, equations involving the summation of forces and moments in the plane of symmetry

Table 13.20 Airspeed indicator correction sequence



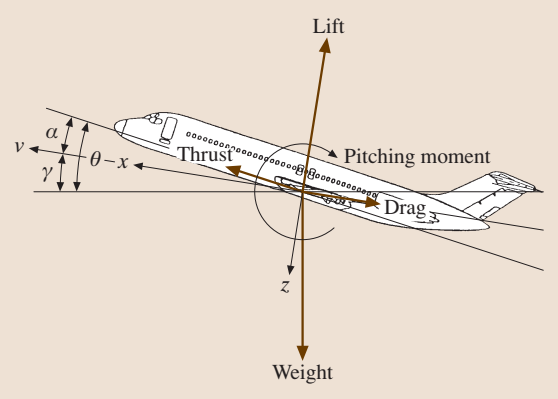


Fig. 13.162 Aerodynamic forces and moments in unaccelerated straight flight

have been developed, so that Newton's laws of motion may be utilized. For unaccelerated symmetric flight along a straight path, the complete forces and moments are as shown in Fig. 13.162.

The summation of the forces and moments for static equilibrium may be written as

$$\sum F_x = T\cos\alpha - D - W\sin\gamma = 0\,,$$
$$\sum F_z = W\cos\gamma - L - T\sin\alpha = 0\,,$$
$$\sum M_{cg} = 0\,.$$

If the assumption is made that the angle of attack is always a relatively small angle, then $\cos\alpha = 1$ and $\sin\alpha = 0$. With this assumption, the force equations reduce to

$$T - D = W\sin\gamma\,,$$
$$L = W\cos\gamma\,. \qquad (13.3)$$

With the small angle assumption of $\sin\gamma = \gamma$, (13.3) may be transformed into an expression for the gradient of climb, that is the gain in altitude over a given distance covered in the horizontal direction.

The climb gradient $= \gamma = (T - D)/W$ and the climb velocity, or rate of climb, is

$$R/C = \frac{(T-D)}{W}V\,. \qquad (13.4)$$

If this is expressed in terms of the lift coefficient, then

$$C_L = \frac{W}{Sq}\cos\gamma$$

or with the small-angle assumption ($\cos\gamma = 1$)

$$C_L = \frac{W}{Sq}\,.$$

13.4.6 Airplane Aerodynamic Characteristics

The concepts associated with the aerodynamic forces and moments acting on the airplane give rise to some important aerodynamic characteristics of the aircraft.

Airplane Lift Curve

In the discussion of aerodynamic forces and moments, it was noted that the lift coefficient is primarily a function of the angle of attack α. The variation of lift coefficient with angle of attack is a very important aerodynamic characteristic of an airplane, and is described in a plot such as that shown in Fig. 13.163, called the airplane lift curve.

It was also noted that the airplane lift coefficient is also dependent on the Mach and the Reynolds numbers, which will be discussed later, but for now, we will focus on the lift curve for a specific Mach number and Reynolds number corresponding to full-scale airplane operation. As seen from Fig. 13.163, the lift curve has a zero value at some, usually negative, angle of attack, a linear region with a well-defined slope $dC_L/d\alpha$, and a departure from the linear slope as the maximum lift coefficient $C_{L_{max}}$ is approached. The characteristics of the lift curve, the zero lift angle, the slope $dC_L/d\alpha$, and $C_{L_{max}}$, depend on certain geometric characteristics of the airplane and its components, as we shall see later. The airplane lift curve has a special relationship to airplane operation in steady, unaccelerated flight. As we have seen for these steady conditions, the variation of the lift coefficient required to balance the weight at various steady flight speeds is as shown in Fig. 13.164. The lowest steady flight speed is called the stalling speed

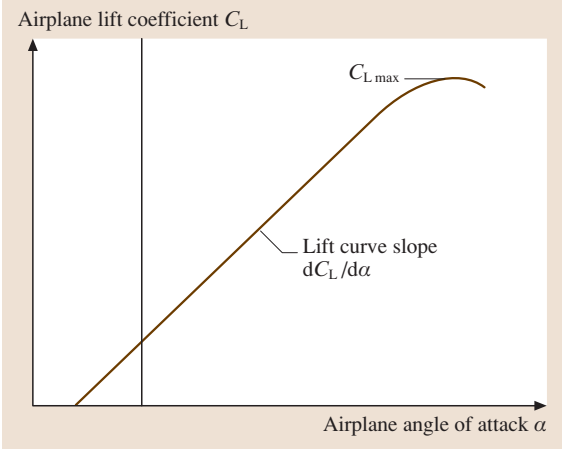


Fig. 13.163 Airplane lift curve

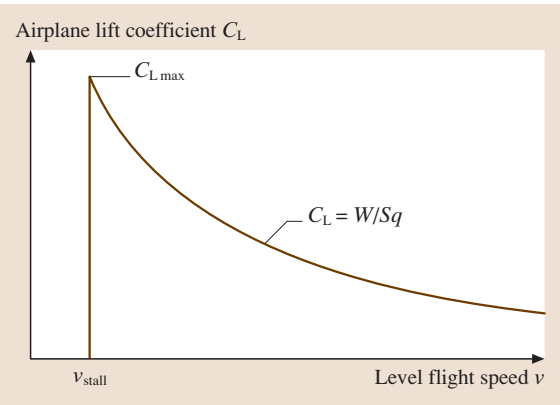


Fig. 13.164 Variation of C_L with airspeed, unaccelerated level flight

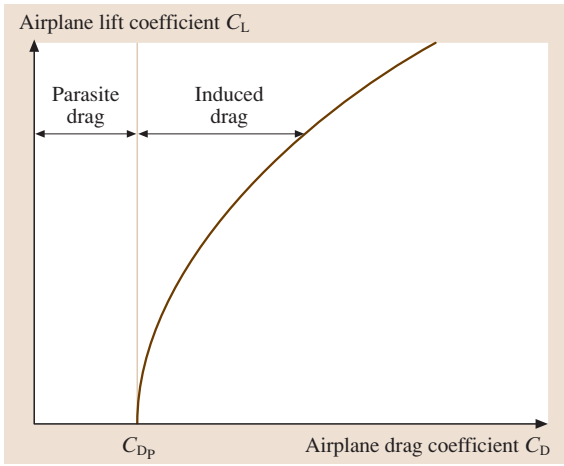


Fig. 13.165 Airplane drag curve

V_{stall} and it corresponds to operation of the airplane at its maximum lift coefficient $C_{L_{max}}$. At high flight speeds, and hence dynamic pressures, the lift coefficients required to balance the weight are reduced as $1/q$ or $1/V^2$. Therefore, the airplane's speed along shallow, unaccelerated flight paths is primarily a function of the lift coefficient, or angle of attack. In order to control the airplane's speed, the pilot must be able to control the equilibrium lift coefficient or angle of attack.

Going back to (13.4), it can be seen that whether the airplane climbs or descends at a given speed depends on the difference between the thrust and drag at this speed. If thrust just equals drag ($T = D$) then the rate of climb will be zero and the airplane will be in level flight. Since the thrust is basically a function of the cockpit throttle setting, under steady unaccelerated flight conditions, the airplane speed is determined by the value of the equilibrium lift coefficient, and the rate of climb or descent is regulated primarily through the throttle. For very large angles of climb or descent, this simple picture does not correspond to actuality, but for the performance methods presented in this chapter, it is a valid concept.

Airplane Drag Curve

Another aerodynamic characteristic of the airplane which is important in range and climb performance is the drag curve or drag polar, a plot of airplane drag coefficient versus airplane lift coefficient. As noted before, the drag varies with angle of attack α but since in the normal operating range of angle of attack, C_L varies linearly with α, it has been found more convenient to describe the drag coefficient as a function of lift coefficient instead of angle of attack. The airplane drag curve, shown schematically in Fig. 13.165, has a value of C_D at zero C_L, called the zero lift or parasite drag coefficient, C_{D_P}. At higher C_Ls, the drag coefficient has a parabolic variation with C_L, due to the induced drag or drag due to lift coefficient, C_{D_i}, which varies as the square of the lift coefficient. As is the case with the lift curve, the drag curve varies in shape with both Mach number and Reynolds number, but for now we will focus on the drag curve that is typical for full-scale airplane operation at one particular Mach number and Reynolds number.

A summary of the physical makeup of airplane drag is presented in Table 13.21

Table 13.21 Physical definition of airplane drag elements

Cruise speed of Ma = 0.5 or less	
Zero lift drag	Skin friction plus pressure drag
Drag due to lift	Subsonic induced drag
Cruise speed between Ma = 0.5 and Ma = 1.0	
Zero lift drag	Skin friction plus pressure drag
Drag due to lift	Subsonic induced drag
Compressibility drag	Drag from local shock waves
Cruise speed greater than Ma = 1.0	
Zero lift drag	Skin friction plus pressure drag
	Supersonic wave drag
Drag due to lift	Supersonic wave drag due to lift
	Subsonic induced drag

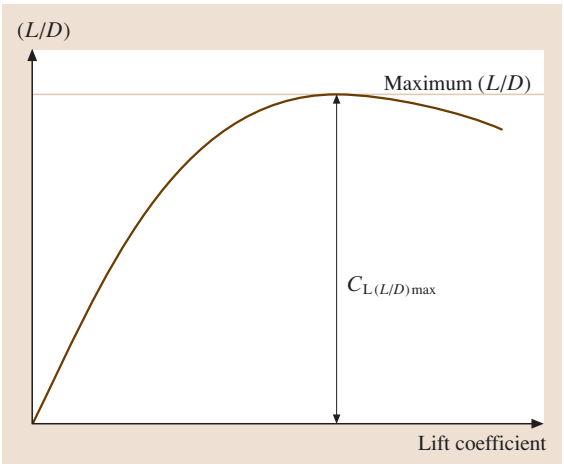


Fig. 13.166 Airplane (L/D) curve

An important parameter in the cruise performance of the airplane, as well as certain aspects of climb performance, is the lift-to-drag ratio L/D. The lift-to-drag ratio may be visualized from the drag polar as shown in Fig. 13.166. At any point on the drag curve, L/D is defined by the ratio C_L/C_D at that point, and also by the slope of a line from the origin to the point in question.

If the L/D values are determined at various points along the airplane drag curve, a plot of airplane L/D versus lift coefficient C_L can be constructed. The L/D value is zero at $C_L = 0$, reaches a maximum value, $(L/D)_{max}$ at some C_L, and then decreases at higher C_L values. Both the value of $(L/D)_{max}$ and the C_L value at which this $(L/D)_{max}$ occurs, $C_{L(L/D)_{max}}$ are important aerodynamic characteristics of the airplane.

Mach Number Effects on Lift and Drag Curves

As noted in the section on flight speed terminology, the characteristics of the airflow around the airplane change dramatically as the flight Mach number is increased due to the compressible nature of air. These changes in the airflow have a significant effect on the airplane lift and drag curves in the various Mach number ranges. For airplanes that operate entirely within the subsonic speed range, there are no significant effects of compressibility of the air on the airplane lift and drag curves, and a single lift curve as shown in Fig. 13.163, and a single drag curve as shown in Fig. 13.165 describe the lift and drag characteristics of the airplane. For airplanes that operate at high subsonic speeds in the transonic speed region, the airplane lift and drag curves will vary as the flight Mach number is increased due to the compressible nature of the air, so that there is a family of lift curves, and a family of drag curves, one for each flight Mach number of interest. The family of lift curves is characterized by an increase in the slope of the lift curve $dC_L/d\alpha$ and a decrease in the maximum lift coefficient $C_{L_{max}}$ as the Mach number is increased in the high subsonic region, as shown schematically in Fig. 13.167.

The Mach number effects may be shown more specifically by plotting the significant parameters as a function of Mach number. For example, the effect of Mach number on increasing the lift curve slope is shown in Fig. 13.168 while the effect of Mach number on de-

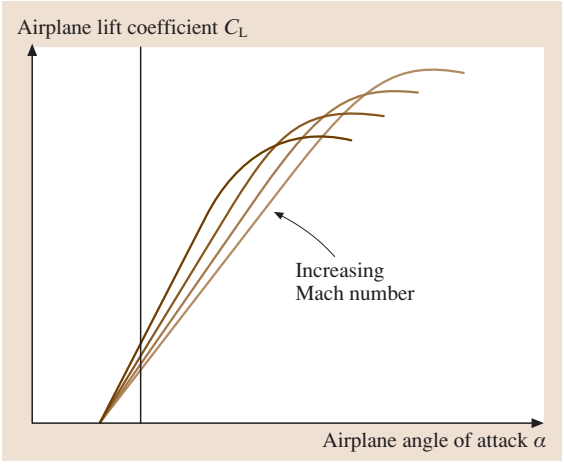


Fig. 13.167 Airplane lift curves at high subsonic Mach numbers

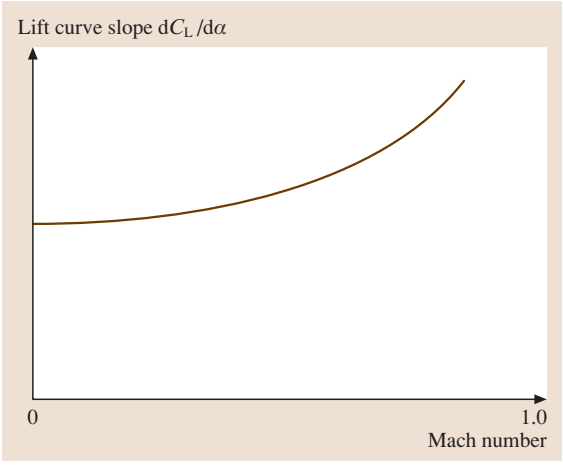


Fig. 13.168 Mach number effect on lift curve slope at high subsonic speeds

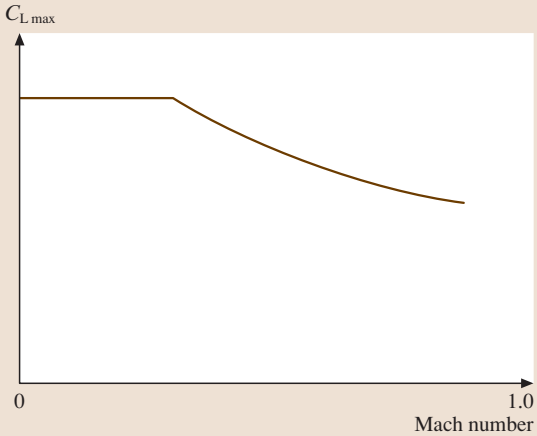


Fig. 13.169 Mach number effect on maximum lift coefficient at high subsonic speeds

creasing the maximum lift coefficient $C_{L_{max}}$ is shown in Fig. 13.169.

The family of drag curves are characterized by increases in the parasite drag coefficient, C_{D_p}, and significant increases in the drag coefficient at higher C_L values as the Mach number is increased, as shown in Fig. 13.170. For the drag curves, the Mach number effects are usually shown in the form of C_D versus Mach number of various values of lift coefficient, as shown in Fig. 13.171.

The explanation of these Mach number effects on the lift and drag curves has been derived from the theory of compressible flow, and confirmed by experimental data obtained in wind tunnels and from flight tests. It

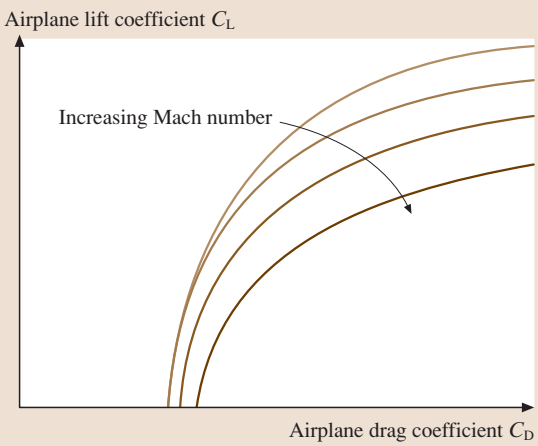


Fig. 13.170 Airplane drag curves at high subsonic speeds

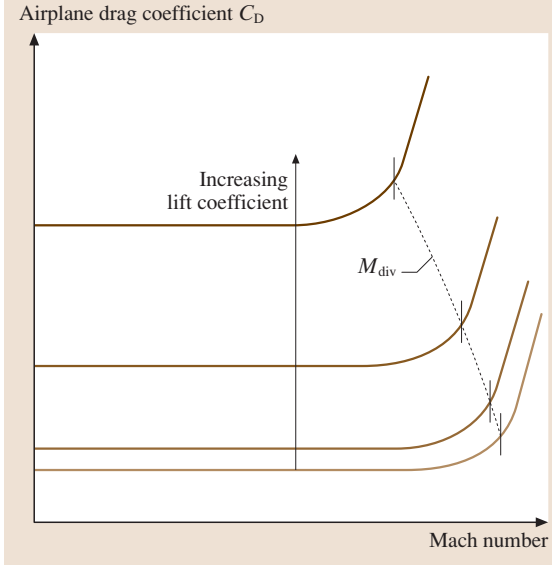


Fig. 13.171 Mach number effect on drag curves at high subsonic speeds

can be shown that, for an airplane at a given angle of attack, the lift coefficient will increase as the Mach number increases, because the suction on the wing upper surface, and the pressures on the wing lower surface tend to grow with Mach number, roughly by the factor $1/\sqrt{1-\mathrm{Ma}^2}$ in the high subsonic speed range, which results in the increase in the lift curve slope as shown in Figs. 13.167 and 13.168. Also, as the Mach number increases, at some flight Mach number, the local velocities on the wing near the leading edge, at high angles of attack near the maximum lift coefficient, become supersonic, which leads to local shock waves and separation, limiting the attainable maximum lift coefficient as shown in Fig. 13.169. As for the drag curves, the Mach number effects are due to the development of local supersonic flow around the wing, which eventually produces normal shock waves, and finally separated flow. Because of the energy loss in the shock wave, and the added pressure drag due to the separated flow, there are significant increases in the drag coefficient at a given lift coefficient as the flight Mach number is increased, as shown in Fig. 13.171. The development of these conditions for a wing airfoil section typical of those used on many current jet transports and business jets is shown in Fig. 13.172. There are some important concepts and definitions associated with the sketches of Fig. 13.172. At the condition shown in Fig. 13.172b, the condition of lift coefficient and flight Mach num-

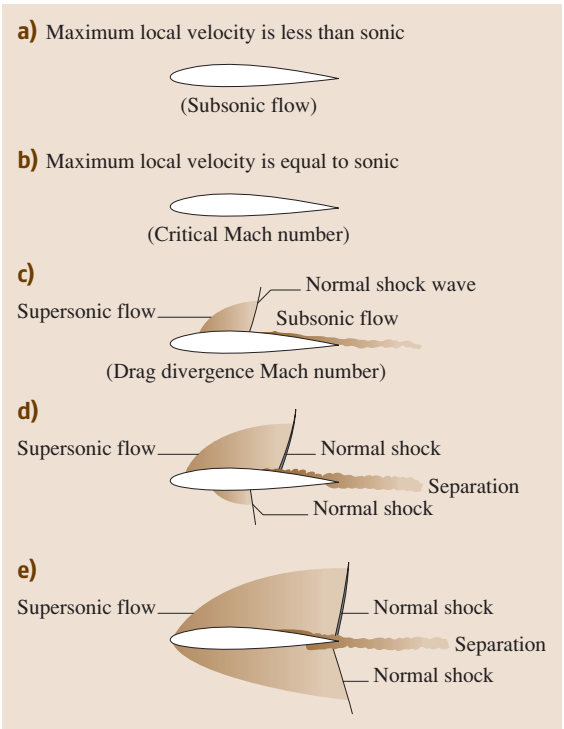


Fig. 13.172a–e Development of flow conditions around a wing airfoil section at high subsonic speeds

ber where the maximum local velocity on the wing surface is equal to the sonic velocity, is called the critical Mach number Ma_C. At a higher Mach number, in the conditions shown in Fig. 13.172c, with a local region of supersonic flow terminated by a normal shock wave, and the very beginning of flow separation behind the normal shock, the drag begins to rise abruptly. This condition is called the drag divergence Mach number Ma_{DIV}, for that particular lift coefficient. At higher flight Mach numbers (Fig. 13.172d,e), the supersonic zones are larger, the normal shocks are stronger, and the drag continues to rise very strongly. Although the wing is the primary source of local supersonic flow, shock waves, and the associated drag increase, all parts of the airplane, i. e., the fuselage, tail surfaces, and engine nacelles will eventually experience these conditions as the flight Mach number approaches 1.0.

For airplanes that are designed to operate at supersonic speeds, the lift and drag curves also vary with flight Mach number. The slopes of the curves are similar

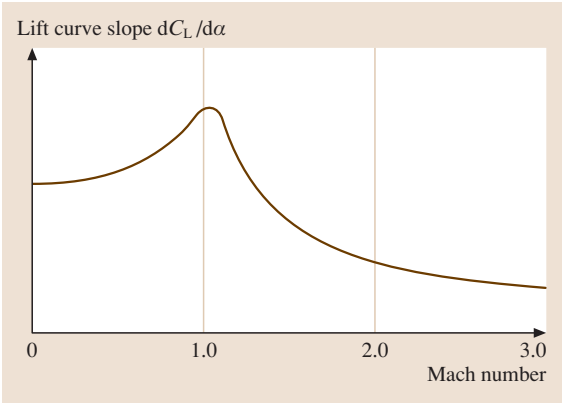


Fig. 13.173 Mach number effects on lift curve slope at subsonic and supersonic speeds

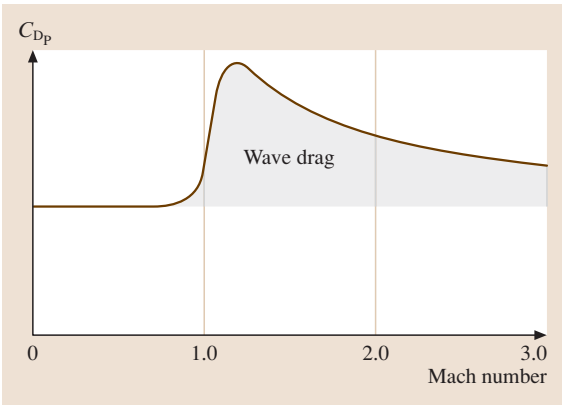


Fig. 13.174 Mach number effects on the parasite drag coefficient

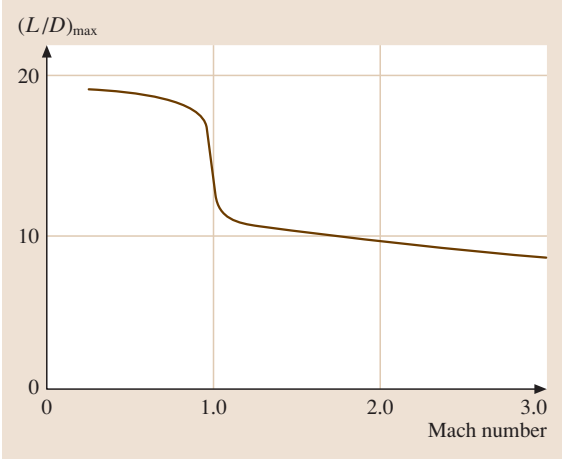


Fig. 13.175 Mach number effects on the maximum lift-to-drag ratio

to those of subsonic speed designs, but the significant parameters show a different trend at supersonic speeds. The pressure on the wing upper and lower surface tend to reduce as the Mach number is increased supersonically, roughly by the factor $1/\sqrt{\mathrm{Ma}^2-1}$, so that the lift curve slope, which increases at high subsonic speeds, decreases beyond Mach 1.0, as shown in Fig. 13.173. The maximum lift capability supersonically is described in terms of a maximum usable lift coefficient, which results from detached shock waves and unsteady flow at high angles of attack. The drag curves at supersonic speeds are basically parabolic in shape, but with very high values of parasite drag coefficient because of the added element of supersonic wave drag, as shown in Fig. 13.174, and very high drag levels at operating lift coefficients due to the wave drag element that increases with lift coefficient. The presence of wave drag at supersonic speeds has a significant impact on the maximum lift-to-drag ratio $(L/D)_{\max}$ referred to in Fig. 13.166. Because of this added drag element, the $(L/D)_{\max}$ values at supersonic speeds are usually less than half of the values at subsonic speeds, for any specific configuration.

For jet transports, the $(L/D)_{\max}$ values subsonically are just under 20, while the best supersonic transport values are less than 10, as shown in Fig. 13.175.

13.4.7 Airplane General Arrangements

Airplanes are often described by their general arrangement. Figure 13.176 shows various aircraft types classified by number of wings, location of the wings on the fuselage, and number of engines.

Aircraft Component Nomenclature

The various components of an airplane have unique names. Figure 13.177 indicates the names of the various components of a modern commercial transport aircraft.

Some specific geometric characteristics are defined in Table 13.22 and shown schematically in Fig. 13.178.

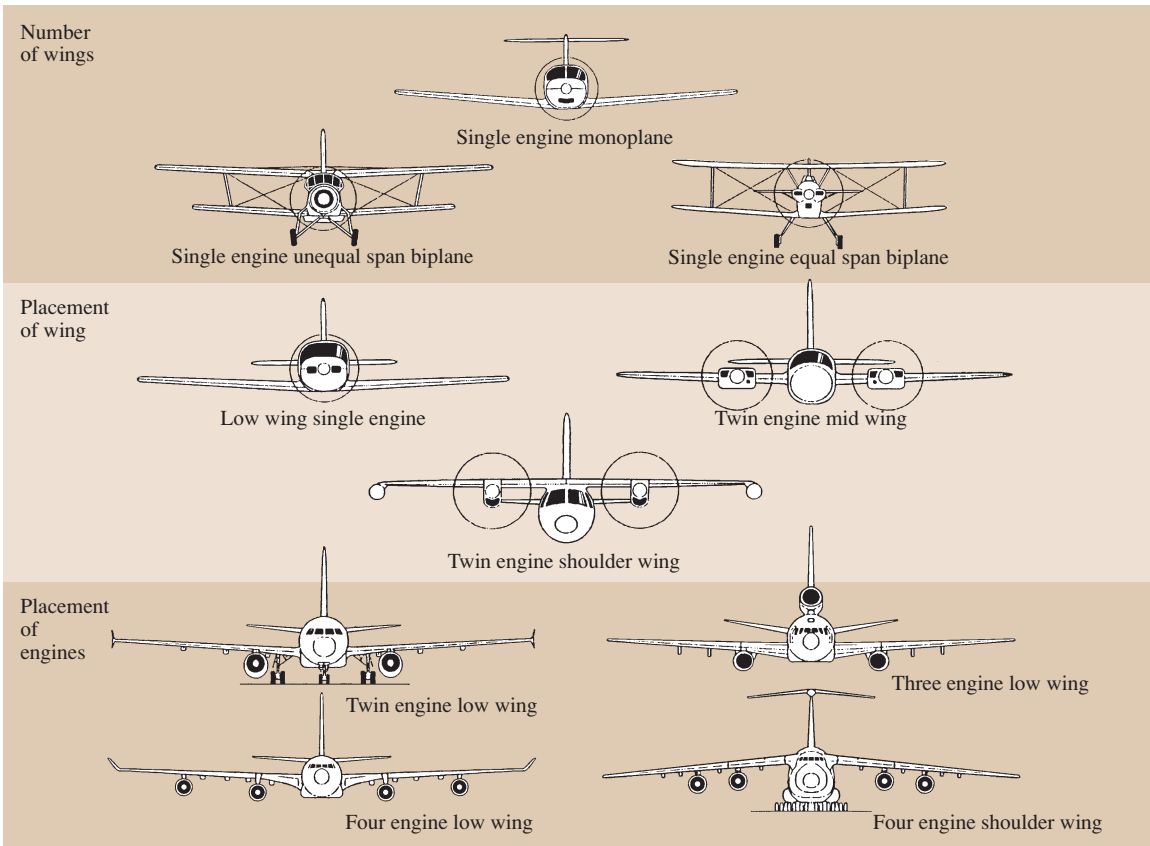


Fig. 13.176 Aircraft general arrangements

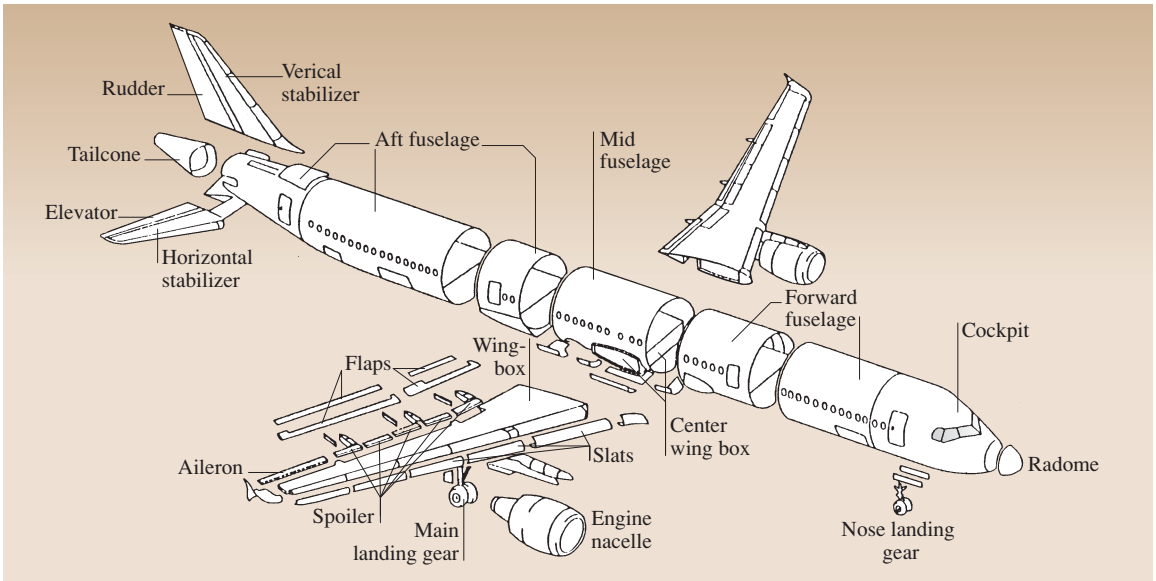


Fig. 13.177 Airplane component nomenclature

Wing Geometric Characteristics

The wing geometric characteristics include the wing area, aspect ratio, taper ratio, airfoil sections, thickness distribution, sweepback angle aerodynamic twist, and dihedral angle. Other important elements that are incorporated in the geometric definition of the wing are high-lift system devices, and lateral control system components.

Table 13.22 Airplane geometric definitions

b	Wing span
$b/2$	Wing semispan
S	Wing area
b_H	Horizontal tail span
S_H	Horizontal tail area
l_H	Horizontal tail length
S_V	Vertical tail area
l_V	Vertical tail length
c	Wing mean aerodynamic chord
c_H	Horizontal tail mean aerodynamic chord
c_V	Vertical tail mean aerodynamic chord
l_{oa}	Overall length
l_{fus}	Fuselage length
h	Height
V_H	Horizontal tail volume
V_V	Vertical tail volume

Wing Area. The wing area is usually taken to be the total planform area of the wing from the fuselage centerline to the wing tip, including the area encompassed by the fuselage, as shown in Fig. 13.178.

Aspect Ratio. The wing aspect ratio (AR) is defined as the square of the wing span b divided by the wing area S

$$\mathrm{AR} = \frac{b^2}{S} .$$

Aspect ratio selection is basically a compromise between aerodynamic efficiency, in the form of high cruise L/D and wing structural weight associated with the bending moments due to air loads for a given wing area.

Taper Ratio. The taper ratio λ is defined as the ratio of the chord at the tip of the wing to the chord at the airplane centerline, called the root chord

$$\lambda = \frac{c_t}{c_r} .$$

Taper ratio is also a compromise between aerodynamic considerations, primarily span load distribution, important for cruise efficiency, stall characteristics, and bending moments due to air loads and structural considerations, associated primarily with the bending moments.

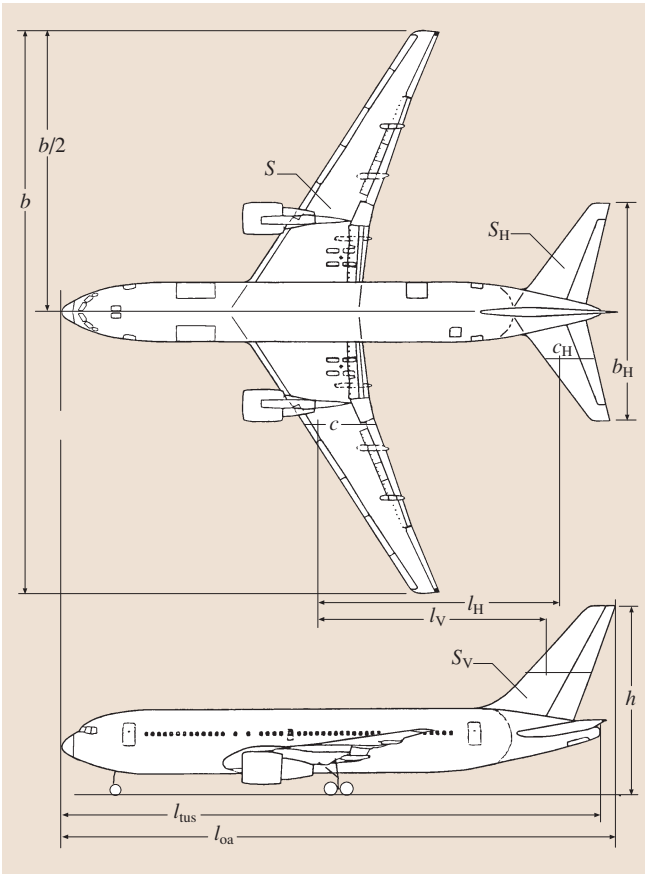


Fig. 13.178 Airplane geometric characteristics

Airfoil Sections. Wing airfoil sections are the cross-sectional shapes of the wing in planes parallel to the airplane center line and normal to the wing reference plan. The wing airfoil sections provide the wing lift, by creating suctions on the wing upper surface and pressures on the wing lower surface. The wing airfoil geometry determines the detailed pressure distribution on the wing upper and lower surfaces, which in turn may have a significant influence on some of the important aerodynamic characteristics of the wing. Airfoil geometric parameters are defined in Fig. 13.179.

Mean Aerodynamic Chord. An important wing geometric parameter is the mean aerodynamic chord (m.a.c.). The m.a.c. is the chord of an imaginary wing, with constant chord, which has the same aerodynamic characteristics as the actual wing. The m.a.c. may be defined graphically, as shown in Fig. 13.180.

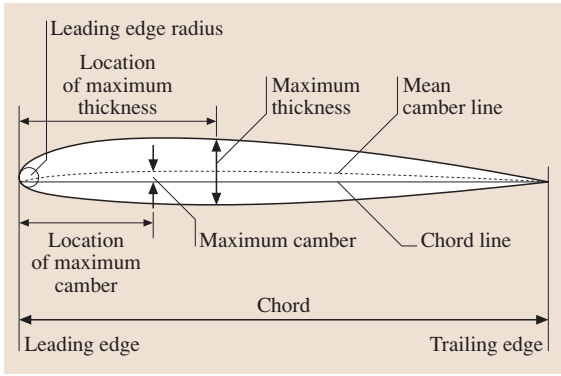


Fig. 13.179 Airfoil section geometric parameters

The wing span b is $b = \sqrt{\mathrm{AR} \cdot S}$, where AR is the wing aspect ratio and S is the reference wing area.

The root chord length is $C_{\mathrm{root}} = \frac{2S}{b(1+\lambda)}$, where λ is the wing taper ratio.

The tip chord length is $C_{\mathrm{tip}} = \lambda C_{\mathrm{root}}$.

For trapezoidal planforms, the wing m.a.c. length is

$$\overline{C} = \left(\frac{2}{3}\right) C_{\mathrm{root}} \frac{1+\lambda+\lambda^2}{1+\lambda} .$$

The distance from the centerline to the m.a.c. location is

$$\overline{Y} = \left(\frac{b}{6}\right) \frac{1+2\lambda}{1+\lambda} .$$

Thickness Distribution. Another wing geometric characteristic is the variation of the airfoil thickness ratio across the span of each wing panel. Of course the simplest wing geometry, still found on many small personal/utility airplanes, is a constant-chord constant-thickness-ratio configuration. More sophisticated personal/utility aircraft, as well as commuters, regional

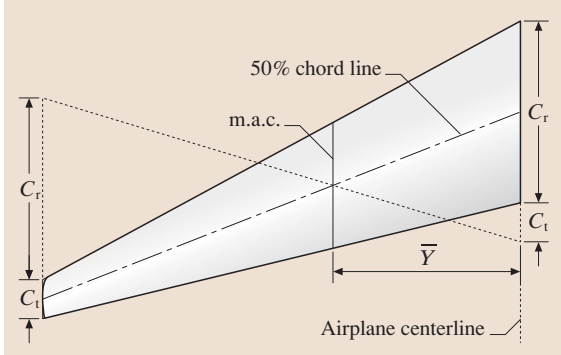


Fig. 13.180 Wing m.a.c. determination

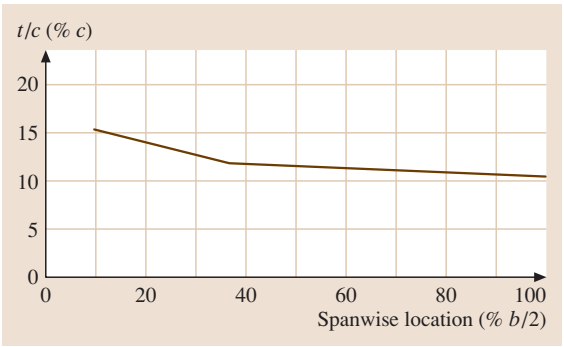


Fig. 13.181 Typical wing thickness distribution for a jet transport

turboprops, business jets, and jet transports have wing designs which vary the airfoil section thickness ratio (t/c) across the span, primarily to obtain greater depth for the airfoil sections at the wing root. This greater depth provides for a more efficient structural beam to resist the bending moments due to wing air loads. For straight-wing propeller-driven aircraft, the wing is usually defined by two airfoil sections, one at the wing root, and one at the wing tip, connected by straight-line elements through the constant percentage chord points.

For business jets and jet transports which cruise at high subsonic speed the wing is usually defined by three or more airfoil sections, one at the side of the fuselage, one at the tip, and one or more at intermediate spanwise locations. The purpose of the additional defining airfoils is to produce wing upper-surface pressure distributions that maintain insofar as possible uniformly swept lines of constant pressure, or isobars at cruise conditions, so that the entire wing reaches its M_{DIV} at the same point. A typical wing thickness distribution for a commercial jet transport is shown in Fig. 13.181.

Sweepback Angle. Two very important wing geometric parameters, especially for airplanes that cruise at high subsonic speeds, are the wing sweepback angle and the average thickness. Wing sweepback or sweep angle is defined in the plan view as the angle between a line perpendicular to the airplane center line and the constant 25% chord line of the wing airfoil sections, as shown in Fig. 13.182. Nearly all high-subsonic-speed aircraft have wings with some amount of sweep, because sweep increases the wing drag divergence Mach number M_{DIV} for a given streamwise airfoil thickness.

Geometric Twist. Nearly all wing designs incorporate some degree of geometric twist, that is, a change in the orientation of the airfoil section chord lines from root to tip, with the tip airfoils having less of an angle of incidence than the root, as shown in Fig. 13.180. This geometric twist is used to avoid initial stalling at the wing tip as the airplane $C_{L_{max}}$ is reached. Typical values of wing twist vary from 3° for personal/utility, commuters, and regional turboprops, to as much as 7° for business jets and jet transports. Supersonic military fighter/attack aircraft usually have little or no twist, and depend on other means to provide satisfactory stall characteristics.

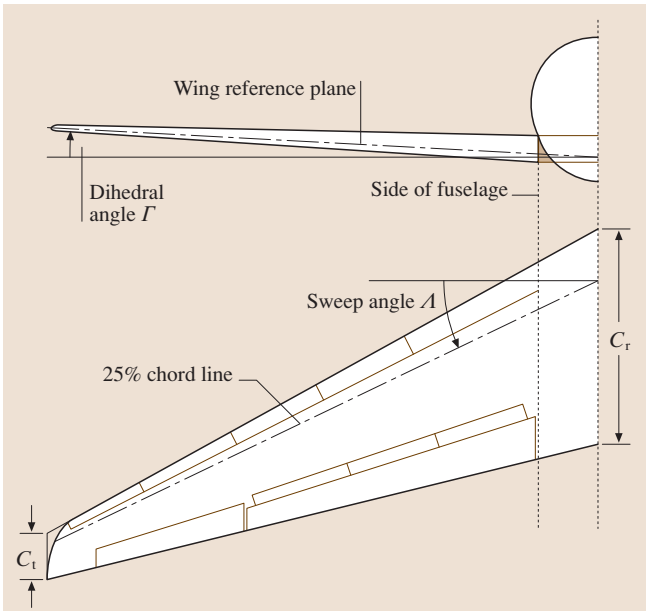


Fig. 13.182 Wing planform parameters

Dihedral Angle. The dihedral angle is defined in the front view of the airplane as the angle between the horizontal and a line midway between the upper and lower surfaces of the wing, as shown in Fig. 13.182. The dihedral angle primarily affects the lateral stability characteristics of the airplane. If the wing tips are higher than the wing root, the wing is said to have positive dihedral, which produces effective lateral stability. For low-wing airplanes with low horizontal tails, nearly all of the lateral stability comes from the wing dihedral angle, which is usually set around 5° for these types. For low-wing airplanes with high-mounted horizontal tails (tee tails), a significant amount of lateral stability is generated by the intersection between the horizontal and vertical tail sections, so that the wing dihedral angle for these types is usually down to around 2°. For

high-wing airplanes, especially those with tee-tail arrangements, there is sufficient lateral stability generated by the wing–fuselage intersection and the horizontal tail–vertical tail intersection that the wing dihedral angle is set at negative angles ranging from −2° to −5°.

Spar Locations. The front and rear spars, along with the upper and lower wing skins, are the major elements of the wing structural *box*, which resists the applied wing loads in bending and torsion. The distance between the wing spars is important structurally, but also has a significant impact on the space available for high-lift and lateral control devices, and on the volume available for internal wing fuel tankage. Typical locations for the front spar are 16–22% of local chord, while typical rear spar locations range from 60% to 75% of local chord.

High-Lift Systems. Nearly all modern aircraft incorporate devices that fit within the wing to increase the maximum lift coefficient in the take-off and landing configuration. These devices are collectively called the high-lift system. High-lift devices fall into two categories: trailing-edge devices located at the rear portion of the wing, and leading-edge devices located at the forward portion of the wing. Trailing-edge devices include single, double, and triple slotted flaps, as shown in Fig. 13.182. Single slotted flaps are standard for personal/utility airplanes. Single slotted flap chords are usually in the range of 25–30% of chord, and have a maximum deflection of 35°. For commuters, regional turboprops, business jets, and jet transports, more powerful double slotted trailing-edge flaps are usually employed. Double slotted flap chords are in the range of 30–35% chord with maximum deflections of 45−50°. Some jet transports with design mission specifications that call for extremely low landing approach speeds and short landing distances have utilized triple slotted trailing-edge flaps to achieve very high $C_{L_{max}}$ values in the landing configuration. Flap chords up to 40% may be used in triple slotted flap designs, with maximum deflections of the aft flap segment of up to 55°. The effectiveness of trailing-edge flaps may be enhanced by selecting a large percentage of chord dimension for the flap, and utilizing as much flap span as possible, considering the need for lateral control ailerons on the outboard wing trailing edge.

Significant increases in $C_{L_{max}}$ can also be achieved through the application of leading-edge high-lift devices. The simplest leading-edge device is the plain leading-edge flap, used on a number of military fighter/attack aircraft. Some jet transports have used leading-edge Krueger flaps. A more effective, but more complicated, leading-edge device is the slat, which is used on all modern jet transports. The maximum effectiveness of leading-edge flaps and slats is usually

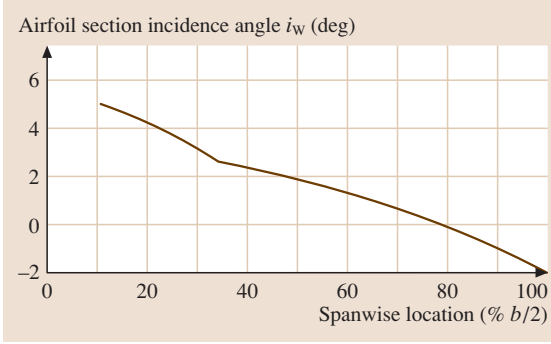


Fig. 13.183 Typical wing twist distribution for a jet transport

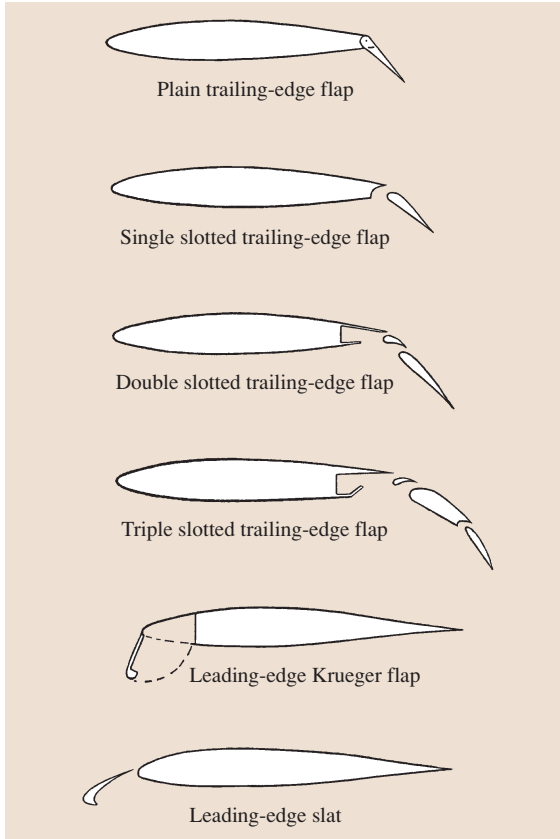


Fig. 13.184 Typical high-lift devices

achieved with flap and slat chords of 12–15%, and deflections ranging from 20° for slats to 30° for plain leading-edge flaps to 60° for Krueger flaps. Leading-edge flaps and slats must extend for the full span of the wing to be effective in increasing $C_{L_{max}}$. Typical high-lift devices are shown in Fig. 13.184.

Lateral Control Devices. An additional consideration for the preliminary wing design is the provision for lateral control devices which produce rolling moments about the airplane's x-axis. The most common lateral control devices are ailerons, essentially a plain trailing-edge flap, and spoilers, basically a portion of the wing upper surface, hinged at its leading edge, which reduces the lift in the affected area of the wing when the spoiler is deflected. When deflected asymmetrically, spoilers can produce significant rolling moments, especially if they are located ahead of deflected trailing-edge flaps. Spoilers have additional uses when deflected symmetrically as drag-producing devices to allow the airplane to slow down in level flight, or to increase the rate of descent at the end of cruise. Spoilers are also used symmetrically to reduce the airplane's lift during landing ground roll, which improves braking effectiveness.

For personal/utility, commuters, and regional turboprops, the usual lateral control device is the aileron, with aileron spans ranging from 55% to 90% of the wing semispan. For business jets and jet transports, spoilers are generally used in conjunction with ailerons. Furthermore, since ailerons located on the outer part of the wing trailing edge tend to twist the wing as they are deflected to produce rolling moments at high dynamic pressure, or high q conditions, thereby reducing there effectiveness, most jet transports utilize smaller ailerons located further inboard in addition to the outboard ailerons for high-q lateral control and trim.

Inboard Trailing-Edge Extensions. Wing inboard trailing-edge extensions are often used on business jets and jet transports with wing sweep angles greater then 30°. The need for an inboard trailing-edge extension arises from the required main landing gear leg upper pivot point location, which would be very near the wing trailing edge on a trapezoidal planform. By incorporating an inboard trailing-edge extension, a suitable structural arrangement may be designed to provide the necessary gear pivot location. This situation is shown in Fig. 13.185.

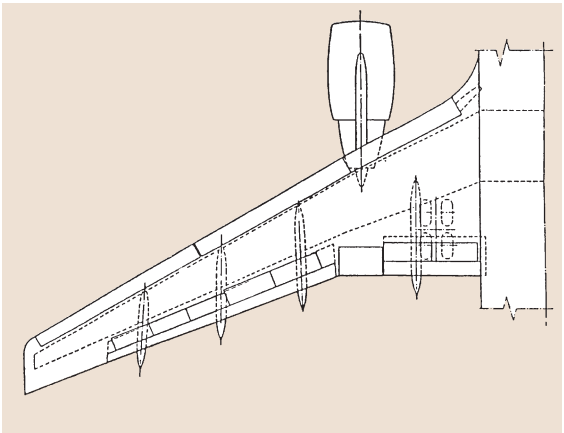


Fig. 13.185 Inboard wing trailing-edge extension

Fuselage Geometry

Primary requirements for the fuselage are to provide accommodation for the crew station, passengers, other payload to be carried in the fuselage, and for some designs, accommodating the engine/propulsion system as well. For single-engine propeller-driven airplanes, the fuselage accommodates the engine/propeller installation forward, the pilot and passenger cabin next, followed by the aft fuselage, which serves mainly as a convenient structure for locating the horizontal and vertical tail well aft of the wing. For larger transport airplanes, the fuselage is made up of three distinct sections: (1) the nose section, (2) a constant-section passenger compartment, and (3) the afterbody or tail cone.

The nose section on larger airplanes usually contains the crew station, with crew seats, control yoke

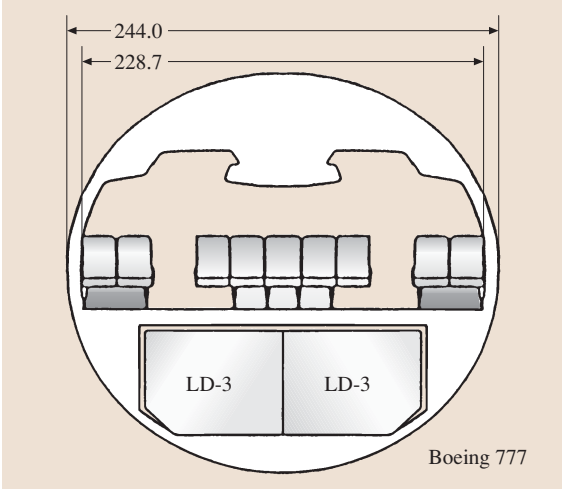


Fig. 13.186 Twin-aisle transport fuselage cross section

and wheel, or control stick, rudder pedals, instrument panel, glare shield, plus a variety of levers, knobs, and switches to operate various aircraft systems. Specific requirements for pilot field of view and downward vision from the defined pilot design eye position are contained in the federal air regulations (FARs) and the military specifications (Mil Specs). The constant-section passenger compartment for larger airplanes is usually pressurized, and circular or near-circular in cross section, because of the structural weight efficiency of this shape for pressure vessels. For larger-capacity long-range jet transports, two aisles are provided for greater passenger mobility, and for ease of entry and exit from multiple adjacent seats. These larger circular cross sections provide significant space below the passenger deck to carry large amounts of revenue cargo, either in special containers or stacked on flat pallets. A typical cross section for a twin-aisle transport is shown in Fig. 13.186.

The length of the passenger compartment must be sufficient to accommodate the required number of passengers, allow for galley space, lavatories, coat rooms, plus passenger entrance doors, and emergency exits.

There are specific requirements in the FARs for emergency exits to be used in survivable accidents. The aft fuselage or afterbody is influenced by conflicting requirements of aerodynamic performance and structural weight. The afterbody should be long enough to avoid severe curvature and separation drag, while being as short as possible to avoid limiting the airplane pitch attitude on the ground during normal take-offs, as well as avoiding excessive weight from a long afterbody.

Empennage Geometry

For conventional aft-tail configurations, the horizontal and vertical tail arrangement, called the empennage, is the major element in providing both static aerodynamic stability in pitch and yaw, as well as providing aerodynamic control moments in pitch and yaw. For unconventional configurations such as *flying wings* or forward horizontal tail *canards*, static aerodynamic stability and control are provided by other means.

The aft horizontal tail is the major contributor to static aerodynamic stability in pitch. This is quite logical, since static longitudinal stability involves the generation of aerodynamic restoring moments which are dependent on an aerodynamic force from the horizontal tail (proportional to the horizontal tail area) and a moment arm (proportional to the distance from the airplane center of gravity to the horizontal tail m.a.c.). The horizontal tail also provides the aerodynamic control moments to allow the pilot to achieve equilibrium in pitch at any desired lift coefficient, allowing the control of airspeed in steady unaccelerated flight, and the curvature of the flight path in accelerated flight. Longitudinal control is usually provided through the hinged, moveable, aft portion of the horizontal tail called the elevator, although some designs move the entire horizontal tail about a fixed pivot point. This arrangement is called an all-moveable horizontal, or a stabilator.

Landing Gear

Most modern airplanes use a tricycle landing gear configuration that is one with two main wheels aft of the c.g. and one forward. Except for small airplanes with low cruise speed, landing gears are usually retracted for climb and cruise flight. The landing gear wheels and tires must be adequate to handle a variety of taxi, take-off, and landing loads prescribed by the FARs, as well as spreading the reaction loads from the gear sufficiently so as not to overstress the runway pavement. The landing gear also has to house the brakes.

Propulsion Systems

Propulsion systems for modern airplanes are of one of the following types:

- Piston engine-propeller
- Turbine engine-propeller
- Turbojet engine
- Turbofan engine
- Turbofan engine with afterburner

Small personal utility or acrobatic airplanes are usually powered by piston engine-propeller combinations, while larger personal utility and smaller commuter airplanes are usually powered by turboprop propulsion. Business jets and larger jet transports are powered by turbofans. High-performance military airplanes are usually powered by low-bypass turbofans equipped with afterburners. The key parameter for the propulsion system is the *specific fuel consumption* c, i. e., the amount of fuel burned per hour, per unit of output of the propulsion system.

For piston engines and turboprops, it is expressed as

$$c = \frac{\text{lb}}{\text{bhp/h}} .$$

For turbojets and turbofans, it is expressed as

$$c = \frac{\text{lb}}{\text{lb/h}} .$$

13.4.8 Weights

The weight of an airplane, especially its empty weight, is a vital parameter in the performance and economics of the design. The following paragraphs provide information on the various aspects of aircraft weight.

Weight Definitions

Figure 13.187 shows a simple bar chart illustrating the elements of the weight buildup to the maximum take-off weight W_{to} required for a specific design mission. Also noted to the right of the bar chart are some important structural weight definitions that are related to the weight buildup, for a typical commercial jet transport. Other aircraft types have similar structural weight definitions.

Note that the operating weight empty involves both the manufacturer's weight empty (MWE), plus the operator's items, which include the flight crew, cabin crew, food, galley service items, drinkable water, cargo containers and pallets, plus life vests, life rafts, emergency transmitters, and the unusable fuel trapped in the fuel system and unavailable for use by the engines. When all the payload is loaded, that is, all available passenger seats filled at the *standard* passenger + baggage weight and all the available revenue cargo volume is filled at some selected cargo density, the aircraft has reached its space limit payload (SLPL), which usually coincides with another key weight definition, the maximum zero-fuel weight (MZFW), the maximum design weight for the aircraft with no fuel on board. Loading on the fuel required for the mission, the mission fuel, plus the required reserves, brings the aircraft to the W_{to}, the maximum take-off gross weight required to perform the specified design mission. The design maximum landing weight (MLW) for most smaller aircraft, such as private propeller-driven aircraft and short-range commuters is usually the same as the MTOGW. However, for larger, long-range aircraft, where the mission fuel is a large percentage of the MTOGW, a somewhat lower MLW is selected to minimize the structural weight impact of designing all the structure to withstand the loads associated with landing at MTOGW. For this design choice, a fuel dump system, which allows fuel to be jettisoned overboard in an emergency following a high-gross-weight take-off, allows the aircraft to reduce its weight to the MLW without having to burn off large amounts of mission fuel. A summary of the airplane weight definitions is given in Table 13.23.

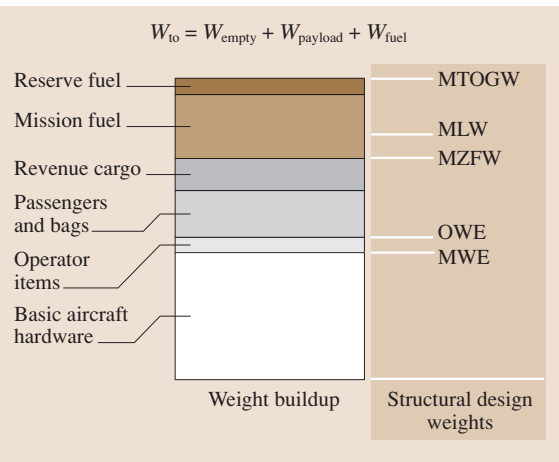


Fig. 13.187 Typical weight buildup – jet transport

Weight Fractions

The maximum take-off weight required for a specific design mission (W_{to}) may be written as

$$W_{to} = W_{empty} + W_{payload} + W_{fuel}$$

or

$$\frac{W_{empty}}{W_{to}} + \frac{W_{payload}}{W_{to}} + \frac{W_{fuel}}{W_{to}} = 1 ,$$

where

$$\frac{W_{empty}}{W_{to}} = \text{weight empty fraction} ,$$
$$\frac{W_{payload}}{W_{to}} = \text{payload fraction} ,$$
$$\frac{W_{fuel}}{W_{to}} = \text{fuel fraction} .$$

and:

W_{empty} = operating weight empty (OWE), the basic aircraft hardware plus other items required to allow the aircraft to perform the design mission;

$W_{payload}$ = passengers + bags + revenue cargo (commercial) or bombs, missiles, cargo (military);

W_{fuel} = total fuel onboard for the mission, that is, fuel burned + reserve fuel.

Weight Estimation and Control

In order to begin design work on a specific aircraft project, there is a need to establish the weight of the various components and systems that make up the aircraft

Table 13.23 Airplane weight definitions

Weight	Symbol	Definition
Manufacturer's weight empty	MWE	Airplane weight at the end of the manufacturing process.
Operating weight empty	OWE	Airplane weight ready for operation. Includes flight crew, cabin crew, food, galley service items, potable water, cargo containers and pallets, life vests, life rafts, emergency transmitter, lavatory fluids, and unusable fuel.
Maximum zero-fuel weight	MZFW	Airplane weight with maximum design payload on board, but no fuel. Design payload includes all passengers and their baggage, plus the maximun design cargo weight.
Maximun landing weight	MLW	Airplane weight defined as the maximum for which the airplane meets all of the structural design requirements for landing. It is usually somewhat higher than the MZFW.
Maximum take-off gross weight	MTOGW	Airplane weight defined as the maximum for which the airplane meets all of the structural design and performance requirements for takeoff. Includes the maximum design payloads plus the fuel required to fly the design mission plus the required reserve fuel.

empty weight. This process, usually conducted by aircraft weight engineers, initially involves much reliance on empirical data from actual aircraft, correlated with appropriate physical parameters. This data is assembled into a group weight statement, a list of weights of the major elements that make up the MWE. Examples are shown in Table 13.24.

As the design work progresses, the group weights are updated as various parts of the aircraft are specifically defined. Major effort is required to ensure that the initial target weight for the entire aircraft is not exceeded during the design and manufacturing phase of the project.

Balance Diagram and C.G. Limits

The airplane balance diagram is used to ensure that the airplane center of gravity (c.g.) is in the proper location for all of the probable loading conditions considering OWE c.g. location, fuel loading and usage, passenger loading, and cargo loading. The OWE c.g. is usually set by designers based on experience for a specific airplane design. Then the extreme excursions of the c.g. due to the most probable adverse loading conditions that move the c.g. forward and aft are examined by calculation to establish forward and aft limits for c.g. travel. The results of these calculations are plotted on a chart of airplane gross weight versus c.g. location so that appropriate c.g. limits can be established. Stability and control and structural design criteria must be met at these c.g. limits.

13.4.9 Aircraft Performance

Aircraft performance is the part of the subject of flight mechanics that deals with parameters such as speed, rate of climb, range, fuel consumption, and runway length requirements.

Level-Flight Performance

The simplest performance condition is steady level-flight cruise, when all forces are in equilibrium as the aircraft moves at a constant speed and altitude. From Fig. 13.162, equilibrium requires that

$$\text{lift } L = \text{weight } W = C_L \frac{\rho}{2} V^2 S \,,$$

$$C_L = \frac{W}{(\rho V^2/2)S}$$

and

$$\text{thrust } T = \text{drag } D = C_D \frac{\rho}{2} V^2 S \quad \text{subsonic flight} \,.$$

For any given weight, C_L may be found and substituted into the induced drag term. ΔC_{D_C}, the empirical compressibility drag coefficient, is dependent on C_L and the Mach number.

Several fundamental airplane characteristics can be derived from the drag equation. One major objective of airplane design is to minimize the drag for any required lift. At any altitude and speed, the ratio of drag to lift depends only on the ratio of C_D to C_L. At low Mach

Table 13.24 Summary of group weight statements. Transport airlines in lb

Weight elements	DC-9-30	737-200	727-100	727-200	707-320	DC-8-55	DC-8-62	DC-10-10	L-1011-1	DC-10-30	747-100
Wing group	11 391	11 164	17 682	18 529	28 647	34 909	36 247	48 990	47 401	57 748	88 741
Tail group	2790	2777	4148	4142	6004	4952	4930	13 657	8570	14 454	11 958
Body group	11 118	11 920	17 589	22 415	22 299	22 246	23 704	44 790	49 432	46 522	68 452
Landing gear	4182	4038	7244	7948	11 216	11 682	11 449	18 581	19 923	25 085	32 220
Nacelle group	1462	1515	2226	2225	3176	4644	6648	8493	8916	9328	10 830
Propulsion group	2190	1721	3052	3022	5306	9410	7840	7673	8279	13 503	9605
Flight controls	1434	2325	2836	2984	2139	2035	2098	5120	5068	5188	6886
Auxiliary power	817	855	0	849	0	0	0	1589	1202	1592	1797
Instruments	575	518	723	827	550	1002	916	1349	1016	1645	1486
Hydraulic system	753	835	1054	1147	1557	2250	1744	4150	4401	4346	5067
Electrical system	1715	2156	2988	2844	3944	2414	2752	5366	5490	5293	5305
Avionics	1108	1100	1844	1896	1815	1870	2058	2827	2801	3186	4134
Furnishings	8594	9119	11 962	14 702	16 875	15 884	15 340	38 072	32 829	33 114	48 007
Air conditioning	1110	1084	1526	1802	1602	2388	2296	2386	3344	2527	3634
Anti-icing system	474	113	639	666	626	794	673	416	296	555	413
Load and handling	57	–	15	19	–	55	54	62	–	62	228
Empty weight (less dry engine)	49 770	51 240	75 528	86 017	105 756	116 535	118 749	203 521	198 968	224 148	297 867
Dry engine	6160	6212	9322	9678	19 420	16 936	17 316	23 229	30 046	25 587	35 700
MEW	55 930	57 452	84 850	95 695	125 176	133 471	136 065	226 750	229 014	249 735	333 567
MTOGW	108 000	104 000	161 000	175 000	312 000	325 000	335 000	430 000	430 000	565 000	775 000

numbers, $\Delta C_{D_C} = 0$ and

$$C_D = C_{D_p} + \frac{C_L^2}{\pi AR e} .$$

For minimum drag, C_D/C_L is a minimum. Now

$$\frac{C_D}{C_L} = \frac{C_{D_p}}{C_L} + \frac{C_L}{\pi AR e} .$$

At the value of C_L for which C_D/C_L is a minimum, $d(C_D/C_L)/dC_L = 0$. Then

$$\frac{d(C_D/C_L)}{dC_L} = -\frac{C_{D_p}}{C_L^2} + \frac{1}{\pi AR e} = 0 .$$

And for $L/D =$ maximum,

$$C_{D_p} = \frac{C_L^2}{\pi AR e} .$$

Thus, for minimum drag, the lift coefficient is the value for which drag due to lift is equal to parasite drag. For this condition

$$C_{L_{(L/D)max}} = \sqrt{C_{D_p} \pi AR e} .$$

The value of L/D is

$$\frac{L}{D} = \frac{C_L}{C_D} = \frac{C_L}{C_{D_p} + C_L^2/(\pi AR e)} ,$$

for $L/D =$ maximum

$$C_L = \sqrt{C_{D_p} \pi AR e}$$

and

$$C_{D_p} = \frac{C_L^2}{\pi AR e} .$$

Thus,

$$\left(\frac{L}{D}\right)_{max} = \frac{\sqrt{C_{D_p} \pi AR e}}{2C_{D_p}}$$

To obtain this minimum drag in flight, we must fly at the speed corresponding to the C_L given above. This speed is designated as $V_{(L/D)max}$. Then

$$V_{(L/D)max} = \sqrt{\frac{2W}{\sqrt{C_{Dp} \pi AR e} \rho S}} .$$

Propeller-driven airplanes achieve their best range at the lift coefficient and corresponding speed for $(L/D)_{max}$.

It is customary to study the performance of propeller-driven airplanes in terms of power, since they operate with engines that produce power rather than thrust. Thrust horsepower required for level flight is drag times distance covered per unit time, so

$$550 \, thp_{req} = DV = C_{D_p} \frac{\rho}{2} V^3 S + \frac{C_L^2}{\pi AR e} \frac{\rho}{2} V^3 S .$$

Then $V = \sqrt{2W/C_L \rho S}$ and we obtain

$$thp_{req} = \frac{1}{550} \sqrt{\frac{2W^3}{\rho S}} \left(\frac{C_{D_r}}{C_L^{3/2}} + \frac{C_L^{1/2}}{\pi AR e} \right) .$$

The constant 550 is carried to keep the units in horsepower and the other units in the corresponding English system units. The minimum power will be obtained when the term in parenthesis is a minimum. Taking the derivative of that term with respect to C_L, equating it to zero, and defining $C_{L_{mp}}$ as the lift coefficient for minimum power required leads to

$$-\frac{3}{2} \frac{C_{D_r}}{C_{L_{mp}}^{5/2}} + \frac{1}{2} \frac{1}{\pi AR e C_{L_{mp}}^{1/2}} = 0 .$$

Therefore

$$C_{L_{mp}}^2 = 3\pi C_{D_p} AR e ,$$

and

$$C_{L_{mp}} = \sqrt{3\pi C_{D_p} AR e} = \sqrt{3C_{L_{(L/D)max}}} .$$

Substituting in the induced drag coefficient portion of the equation gives

$$C_{D_{imp}} = \frac{3\pi C_{D_p} AR e}{\pi AR e} = 3C_{D_p} .$$

At the minimum-power condition, the induced drag coefficient is three times as large as the parasite drag coefficient. This contrasts with the minimum drag condition, for which they are equal. Since, for a given total lift, the speed varies inversely as the square root of the lift coefficient, the speed for minimum power is lower than the speed for minimum drag by the ratio $1/(3)^{1/4} = 0.76$. Taking the inverse, the minimum drag speed is 1.32 times the minimum-power speed.

Climb and Descent Performance

Figure 13.162 illustrates the force on an airplane in steady-state constant-speed climb. The thrust is shown acting parallel to the flight path direction. In general, this is not quite true, but in conventional aircraft the effects of an inclination of the thrust vector are small enough to be neglected.

Equating forces perpendicular and parallel to the flight path

$$L = W \cos \gamma ,$$
$$T = D + W \sin \gamma .$$

Then

$$\sin \gamma = \frac{T - D}{W} = \frac{T}{W} - \frac{D}{W} = \frac{T}{W} - \frac{D}{L} ,$$

γ is the flight path angle or angle of climb. We may assume that γ is sufficiently small so that $\cos \gamma$ is approximately equal to 1.0. Then $L = W$.

$$\text{rate of climb RC} = V \sin \gamma = \frac{V(T - D)}{W}$$

For propeller-driven aircraft, it is convenient to use power rather than thrust and drag. If RC is to be determined in feet per minute, the usual units, then with V

in feet per second,

$$\begin{aligned}\text{RC(ft/min)} &= 60\left(\frac{TV - DV}{W}\right)\\ &= \frac{\text{thp}_{\text{avail}} - \text{thp}_{\text{req}}}{W}(33\,000)\\ &= \frac{\text{thp}_{\text{excess}}}{W}(33\,000)\,,\end{aligned}$$

where $\text{thp}_{\text{excess}}$ is the thrust horsepower available for climbing; W is in pounds.

The preceding equations are based on an airplane climbing at constant true airspeed. In practical operations, climbing flight is done at constant indicated airspeed or constant Mach number. This provides the pilot with a simple guide to the proper climb speed, whereas a constant true speed would mean an ever-changing indicator reading as the altitude increases. A constant indicated speed essentially corresponds to a constant calibrated airspeed. At low Mach numbers, this is the same as a constant equivalent airspeed, since the compressibility correction to airspeed is very small.

With a constant equivalent airspeed, the airplane continually accelerates as the altitude increases. The equilibrium equation along the flight path must then include an inertial term. Thus

$$T = D + W\sin\gamma + \frac{W}{g}\frac{\mathrm{d}V}{\mathrm{d}t}\,.$$

Since

$$\frac{\mathrm{d}V}{\mathrm{d}t} = \frac{\mathrm{d}V}{\mathrm{d}h}\frac{\mathrm{d}h}{\mathrm{d}t} \text{ and } \frac{\mathrm{d}h}{\mathrm{d}t} = V\sin\gamma$$

we can write

$$\sin\gamma = \frac{(T-D)/W}{1+(V/g)(\mathrm{d}V/\mathrm{d}h)} = \frac{T-D}{W}\,\text{(K.E. factor)}\,.$$

This equation differs from the previous rate-of-climb equation at constant true speed by the kinetic-energy correction factor $[1+(V/g)(\mathrm{d}V/\mathrm{d}h)]^{-1}$. Approximate values of the term $(V/g)(\mathrm{d}V/\mathrm{d}h)$ are given as functions of Mach number in Table 13.25 for various types of climb paths. Note that, for constant-Mach-number climb below the isothermal atmosphere, the correction increases the rate of climb. In this region, constant Mach number means decreasing velocity as altitude increases because the speed of sound is decreasing. The airplane is losing kinetic energy and trading it for increased rate of climb. When applicable, the kinetic-energy correction factor is applied to gradient-of-climb and rate-of-climb calculations.

The foregoing equations include a surprisingly large amount of useful information. First, the federal air regulations (FARs) specify minimum permissible performance for commercial aircraft in terms of minimum climb gradients, primarily after failure of one engine. A gradient is the tangent of an angle. For small to moderate flight path angles, the sine of the angle is essentially equal to the tangent, so

$$\begin{aligned}\text{gradient } \gamma &= \tan\gamma = \sin\gamma\\ &= \frac{T-D}{W}\,\text{(K.E. factor)}\\ &= \left(\frac{T}{W} - \frac{D}{W}\right)\text{(K.E. factor)}\,.\end{aligned}$$

Thus the gradient depends on the thrust-to-weight ratio minus the inverse of the lift-to-drag ratio.

There are three conditions that are related to an airplane's climb performance. The first is the gradient of climb, important for clearing obstacles close to the ground. The second is the rate of climb, important for

Table 13.25 Kinetic-energy correction factors for climb

Climb operation	Altitude (ft)	$\frac{V}{g}\frac{\mathrm{d}V}{\mathrm{d}h}$ (approx.)
Constant true speed	All	0
Constant v_E	Above 36 150	0.7Ma^2
Constant v_E	Below 36 150	0.567Ma^2
Constant Ma	Above 36 150	0
Constant Ma	Below 36 150	-0.133Ma^2

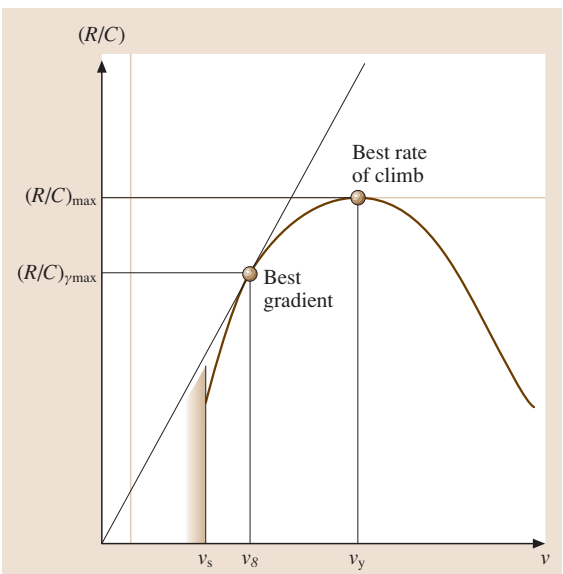


Fig. 13.188 Climb speed diagram

reaching cruise altitude as soon as possible. The third is the best economy of climb, climbing at a higher speed than the best rate of climb speed, but covering more distance in the climb for a given amount of fuel burned. These climb speeds are shown in Fig. 13.188.

Other important performance characteristics can be found from the climb gradient and rate-of-climb equations. If the thrust is zero, then the flight path gradient is the inverse of the L/D ratio, and the rate of descent or sink speed is given by the expression

$$\text{rate of descent} = V/(L/D) .$$

For gliders and sailplanes, these equations are used to develop charts such as the one shown in Fig. 13.189, which provide the essential performance data for these aircraft.

Range

The total range capability of an airplane consists of the climb, cruise, and descent segments. The cruise range is equal to the summation of the range increments obtained by multiplying the miles flown per pound of fuel used (mi/lb) at selected average weights by the appropriate incremental fuel quantity. The value of the miles per pound is called the specific range and is defined as follows.

For jet or turbofan aircraft,

$$\frac{\text{mi}}{\text{lb}} = \frac{\text{miles flown per hour}}{\text{fuel flow (lb/h)}} = \frac{V}{cT} = \frac{V}{cD} ,$$

where c is the specific fuel consumption (SFC) in pounds of fuel per pound of thrust per hour (lb/(lb/h)).

The SFC used in the performance equations is the installed specific fuel consumption. *Installed* means that all adverse effects on SFC associated with the engine installation are included. A typical markup on engine specification SFC for jet and turbofan transports is 3% (i. e., installed SFC equals 1.03 times bare engine SFC). In some cases, for which the engine data are based on zero inlet and nozzle losses, the markup may be twice that amount.

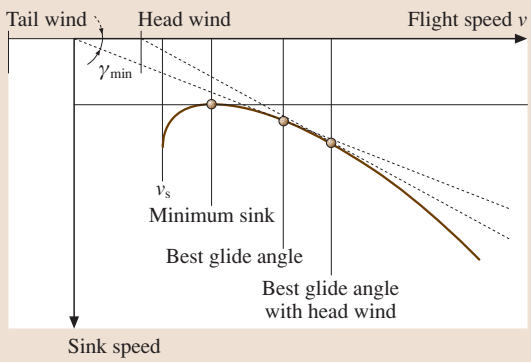


Fig. 13.189 Glider performance chart

Also,

$$D = \frac{W}{L/D} = \frac{D}{L} W$$

and thus

$$\text{mi/lb} = \frac{V}{c(D/L)W} = \frac{V}{c}\frac{L}{D}\frac{1}{W} \text{ (jet)} .$$

The term $(V/c)(L/D)$ is called the range factor and is a measure of the aerodynamic and propulsive system range efficiency. If V is in knots, the range in miles per pound will be in nautical miles per pound of fuel.

For a propeller-driven airplane,

$$\frac{\text{mi}}{\text{lb}} = \frac{V}{c\text{bhp}} ,$$

where c is the specific fuel consumption in pounds of fuel per horsepower per hour (lb/(bhp/h)). Then the range in nautical miles per pound of fuel is

$$\frac{\text{mi}}{\text{lb}} = \frac{V(\text{kn})}{c(\text{thp}/\eta)} = \frac{\eta}{c}\left(\frac{V(550)}{DV \times 1.69}\right) = 325\frac{\eta}{c}\frac{1}{D} = 325\frac{\eta}{c}\frac{L}{D}\frac{1}{W} \text{ (propeller)} .$$

For propeller-driven airplanes, specific range depends only on propeller efficiency, SFC, and L/D; speed is not a factor. Since propeller efficiency and SFC are essentially constant in cruise, range is determined by L/D.

The specific range equation for jet airplanes can also be written as

$$\text{mi/lb} = a\frac{\text{Ma}}{c}\frac{L}{D}\frac{1}{W} ,$$

where a is the speed of sound and Ma is the cruise Mach number. The expression $(\text{Ma}L/D)$ is a measure of the specific range capability due to the aerodynamic characteristics of jet powered airplanes. Curves of $(\text{Ma}L/D)$ versus lift coefficient at various Mach numbers show that jet range is increased by high speed as well as high L/D up to the point where the adverse compressibility drag effect on L/D overpowers the beneficial effect of higher speed. This situation can be summarized on a chart such as that in Fig. 13.190, which shows that,

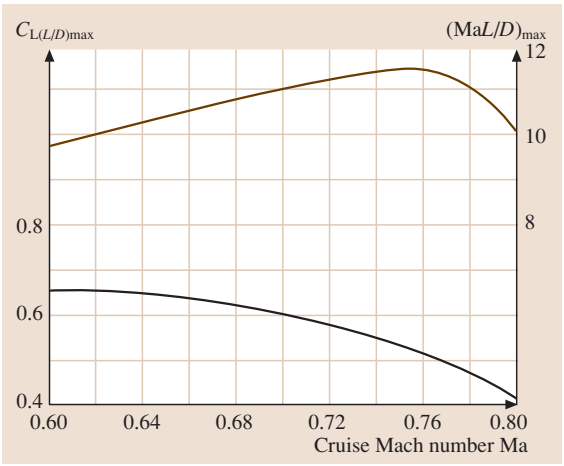


Fig. 13.190 $(MaL/D)_{max}$ and $C_{L(L/D)max}$ for maximum range, jet airplanes

if the SFC is constant with Mach number, the maximum mi/lb value for a jet airplane occurs at a Mach number and lift coefficient where the maximum value of (MaL/D) at each Mach number reaches its peak. However, the turbofan SFC increases slowly with Mach number, so that the optimum range will be obtained at a slightly lower Mach number than taht indicated by Fig. 13.190.

The total cruise range is given by

$$\text{range} = \int_{w_f}^{w_i} \frac{\text{mi}}{\text{lb}}\,\mathrm{d}W\,.$$

If average values of η, c, and L/D can be chosen, the cruise range may be found analytically. For jets,

$$\begin{aligned} R(\text{n. miles}) &= \int_{w_f}^{w_i} \frac{\text{mi}}{\text{lb}}\,\mathrm{d}W \\ &= \int_{w_f}^{w_i} \frac{V}{c}\frac{L}{D}\frac{\mathrm{d}W}{W} \\ &= \frac{V}{c}\frac{L}{D}\log_e\frac{W_i}{W_f}\,. \end{aligned}$$

For propeller-driven planes,

$$\begin{aligned} R(\text{n. miles}) &= \int_{w_f}^{w_i} 325\frac{\eta}{c}\frac{L}{D}\frac{\mathrm{d}W}{W} \\ &= 325\frac{\eta}{c}\frac{L}{D}\ln\frac{W_i}{W_f} \quad \text{(Breguet formula)}\,, \end{aligned}$$

where η, c, and L/D are assumed constant throughout the flight or, more realistically, are taken as effective averages; V is in knots; and c is in lb/(lb/h) or lb/(bhp/h), as appropriate. The range formula for propeller-driven aircraft is called the Breguet formula, after its originator. The analogous jet formula is often similarly labeled.

The total range includes the cruise range plus the distance covered in climb and descent. For long flights, where the cruise portion is dominant, the lower mi/lb in climb due to the higher power being used, and the higher mi/lb in descent due to lower power being used, may be assumed to cancel, allowing the range to be estimated directly from the Breguet range equation, using the appropriate parameters chosen at an average weight. For short flights, where climb is a large portion of the total trip, this approximation may result in a significant error.

A summary of the payload–range characteristics of an airplane is shown in Fig. 13.191. The payload–range curve is one of the most important performance curves for a commercial transport or business jet. It establishes the envelope which shows how far the airplane can carry a given payload, or how much payload it can carry over a given range.

The payload–range curve applies to a specific airplane–engine combination (MWE, OWE, interior seating arrangement, fuel-tank configuration, MTOGW), operating under specific flight rules (cruise Mach number, altitude, altitude steps, alternate distance). The maximum payload is usually the volume or space limit payload (full passengers + bags and full cargo containers or pallets at some standard cargo density, i. e., 10 lb/ft^3) or the maximum zero-fuel weight limit payload, based on the structural limit of maximum zero-fuel weight. When operating on the maximum

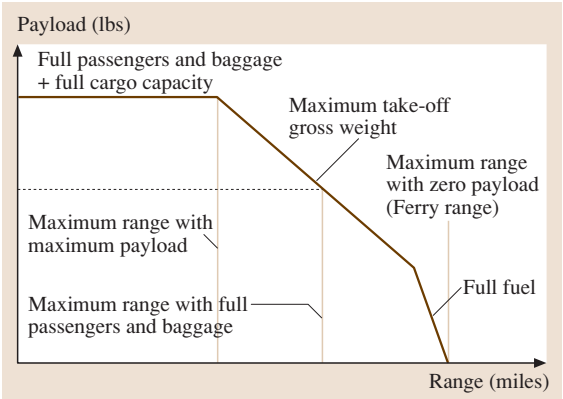


Fig. 13.191 Payload–range curve

take-off weight limit, it is necessary to trade payload for fuel if greater range is desired. Operating on the fuel capacity limit line requires large reductions in payload to achieve small increases in range, due to modest improvements in cruise efficiency achieved by reductions in cruise weight. The key points on the payload range curve are calculated using the Breguet range equation for jet aircraft

$$\text{Range (n. miles)} = \left(\frac{V}{c}\right)\left(\frac{L}{D}\right)\ln\frac{W_\text{i}}{W_\text{f}}\,.$$

For a specific design, on a particular cruise operation, V, C, and L/D are usually taken as constants, W_initial and W_final are derived from known weights, i. e., OEW, payload, reserve fuel, maximum take-off weight, maximum fuel capacity. From this equation, it can be seen that maximum range is achieved by cruising at a point where the quantity VL/D is a maximum.

Some key ideas about payload range curves are as follows:

- Weight-limited payload = MZFW − OWE.
- Space-limited payload is slightly lower.
- Passengers + bags range is usually set by MTOGW.
- Greater passengers + bags range can be achieved by increasing MTOGW (allows more fuel to be carried).
- If passenger + bags range is limited by maximum fuel capacity, greater range can be achieved only through increased fuel capacity.

Endurance

The endurance problem is similar to the range problem except that we are trying to determine how long the aircraft will fly rather than how far it will fly. The quantity analogous to specific range is specific endurance, the hours flown per unit quantity of fuel. In the usual units, specific endurance is measured in hours per pound of fuel.

To maximize endurance, fuel flow per unit time must be minimized. Since the specific fuel consumption is nearly constant, the drag must be minimized for jet aircraft, while thrust horsepower must be as small as possible for propeller-driven aircraft. The endurance t_e for turbojet or turbofan aircraft is

$$t_\text{e} = \int_{w_\text{f}}^{w_\text{i}} \text{h/lb fuel}\,\mathrm{d}W = \int_{w_\text{f}}^{w_\text{i}} \frac{1}{Dc}\,\mathrm{d}W$$

$$= \int_{w_\text{f}}^{w_\text{i}} \frac{1}{[W/(L/D)]c}\,\mathrm{d}W$$

$$= \int_{w_\text{f}}^{w_\text{i}} \frac{1}{c}\frac{L}{D}\frac{\mathrm{d}W}{W} = \frac{1}{c}\frac{L}{D}\ln\frac{W_\text{i}}{W_\text{f}}\,.$$

Again, c and L/D are assumed to be constant throughout the flight, or taken as average values. Note the similarity between this equation and the jet range equation.

Endurance is simply range divided by speed. For the greatest endurance, the aircraft should obviously fly at the speed for minimum drag. This, of course, assumes that c is a constant with speed, a good assumption for jets but not quite true for turbofans. In addition, at very low engine thrust levels, c tends to increase as the thrust is decreased. This may also influence the speed for best endurance.

For propeller-driven aircraft, endurance is

$$t_\text{e} = \int_{w_\text{f}}^{w_\text{i}} \frac{1}{\text{thp}\,c/\eta}\,\mathrm{d}W = \int_{w_\text{f}}^{w_\text{i}} \frac{\eta}{c}\left(\frac{550}{DV\times 1.69}\right)\mathrm{d}W$$

$$= \int_{w_\text{f}}^{w_\text{i}} 325\frac{\eta}{c}\frac{L}{DV}\frac{1}{W}\,\mathrm{d}W\,,$$

where V is in knots and c is in pounds of fuel per brake horsepower per hour. We cannot assume that L/DV is a constant. L/DV is the ratio of lift to thrust power required, and we have shown that thp_req is a nonlinear function of weight. At any given lift coefficient, power required varies with $W^{3/2}$. However, it has been shown that, if V is expressed as $\sqrt{2W/C_\text{L}\sigma\rho_\text{s}S}$, t_e is given by

$$t_\text{e}(\text{h}) = 37.9\left(\frac{\eta}{c}\right)\frac{C_\text{L}^{3/2}}{C_\text{D}}\sqrt{\frac{\sigma S}{W_\text{i}}}\left[\left(\frac{W_\text{i}}{W_\text{f}}\right)^{1/2} - 1\right].$$

Here we see that, for best endurance, a propeller-driven airplane should be flown at the flight condition for maximum $C_\text{L}^{3/2}/C_\text{D}$. It is a maximum when the lift coefficient is $\sqrt{3}$ times the value for maximum lift-to-drag ratio. Since low speed as well as a particular lift coefficient is desirable for minimum power required, best endurance occurs at a high density (i. e., low altitude). Therefore, propeller aircraft endurance is best at low altitudes.

Take-Off Performance

The take-off performance problem is basically an acceleration to the required speed plus a transition to climb

at a 35 ft height for civil turbine-powered transports, or a 50 ft height for piston-engine general aviation or military airplanes. The required runway length is defined as the distance from the start of take-off to the point where these *obstacle* heights are reached. The take-off field length required by FARs for jet transport operation is the greater of:

1. The all-engine take-off distance ×1.15
2. The take-off distance with an engine failure at the *most critical point* in the take-off

The most critical point is where the distance to stop on the runway is equal to the distance to continue the take-off with one engine failed to a height of 35 ft. This situation is called the balanced field length concept. The most critical point in a given take-off, where the *accelerate–stop* distance is equal to the *accelerate–continue* distance is found by plotting the accelerate–stop distance and accelerate–continue distance versus the engine failure speed, as shown in Fig. 13.192.

For propeller-driven transports, the FAR balanced field length concept is the same, but the speed at 35 ft is $1.15V_S$ for four or more engines, and $1.2V_S$ for two or three engines.

For single and multi-engine normal, utility, and acrobatic aircraft under 12 500 lb maximum gross weight, FAR 23 specifies only the all-engine take-off distance to a height of 50 ft with a speed equal to or greater than $1.3V_S$ at 50 ft. For FAR 23 commuter-category aircraft, the balanced field length concept is applied and the criteria for take-off are basically the same as for FAR 25, except that the take-off distance is defined to a 50 ft height, and the speed at 50 ft must be at least $1.3V_S$.

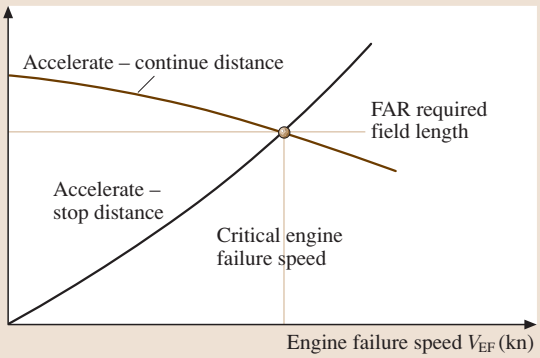


Fig. 13.192 Determination of FAR balanced field length with engine failure

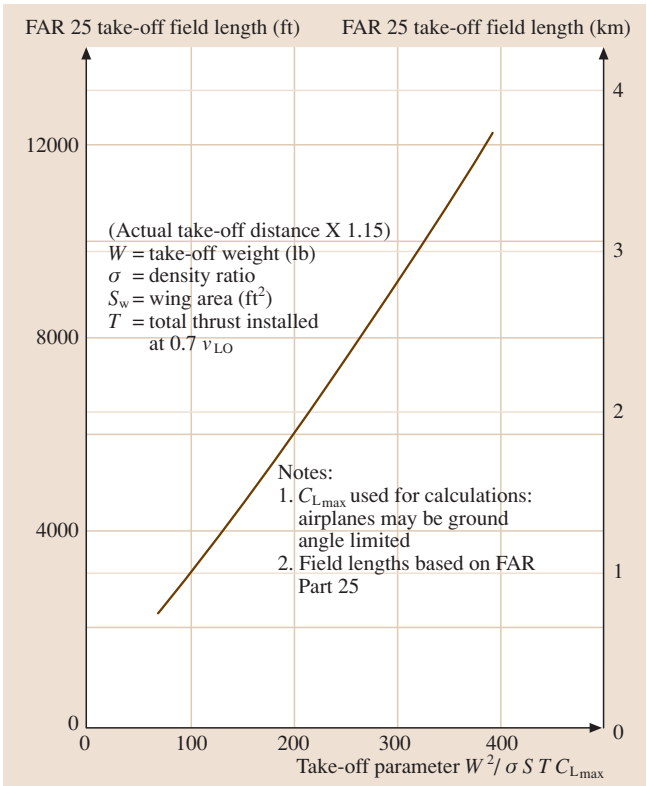


Fig. 13.193 FAR 25 all-engines-operating take-off field length to a 35 ft height

Analysis and flight test suggest that the distance to liftoff is a function of several airplane factors, namely

$$d_{LO} = f\left(\frac{W^2}{\sigma S C_{L_{max}} T}\right),$$

where $\sigma = \rho/\rho_s$ and T is the thrust at $0.7V_{LO}$. Since the ground run distance to lift-off is about 80% of the total distance to the 35 ft height, it has been shown that this parameter works very well in correlating the required runway length results for many aircraft. Figure 13.193 shows the FAR all-engines-operating take-off field length to a 35 ft height d_{35} for jet or turbofan aircraft as a function of $W^2/\sigma S C_{L_{max}} T_{0.7V_{LO}}$. This chart applies to normal take-off without engine failure and includes a 15% increase above the actual performance in accordance with the air transport requirements of FAR 25.

The actual distance is the distance determined from Fig. 13.193 divided by 1.15. Figure 13.194 shows the FAR one-engine-inoperative take-off field length for two-, three-, and four-engine transports.

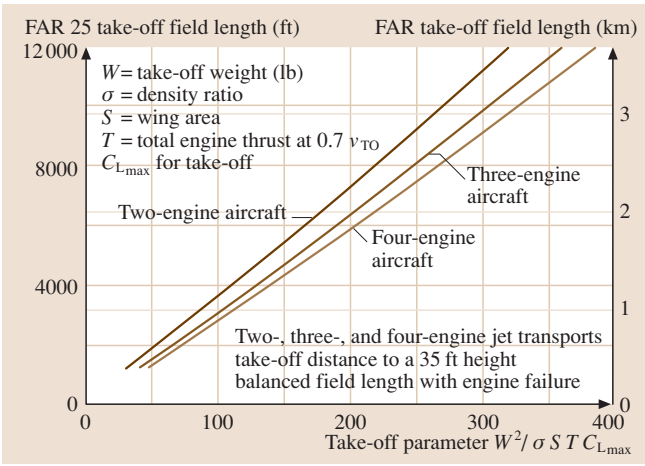


Fig. 13.194 FAR 25 one-engine-inoperative take-off field length

For propeller-driven aircraft, a comparable analysis shows that take-off distance is a function of $W^2/\sigma S C_{L_{max}} P$, where P is the total brake horsepower. This is a less accurate approximation, since the effectiveness of the power depends on the propeller efficiency during take-off. Use of this parameter assumes that all propellers are designed to attain a similar level of efficiency in take-off.

Landing Performance

Landing distances consist basically of two segments: the air run from a height of 50 ft to the surface accompanied by a slight deceleration and flare, and the ground deceleration from the touchdown speed to a stop. Landing runway lengths are required by FAR 25 for commercial aircraft to be demonstrated by flight tests. The air distance d_{air} can be approximated by a steady-state glide distance d_{GL} plus an air deceleration distance d_{decel} at constant altitude, as shown in Fig. 13.195.

V_{50} is the speed at the 50 ft height. In accordance with FAR 25, V_{50} must be at least $1.3V_S$. In practice, it is taken as equal to $1.3V_S$. V_L is the landing or touchdown

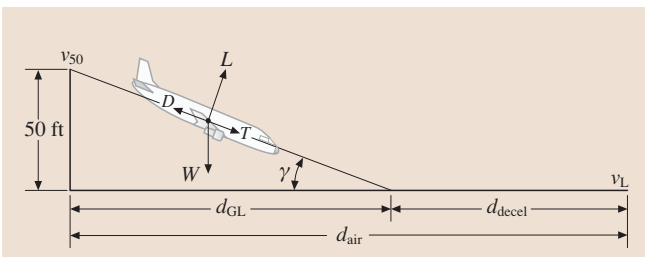


Fig. 13.195 Landing air distance

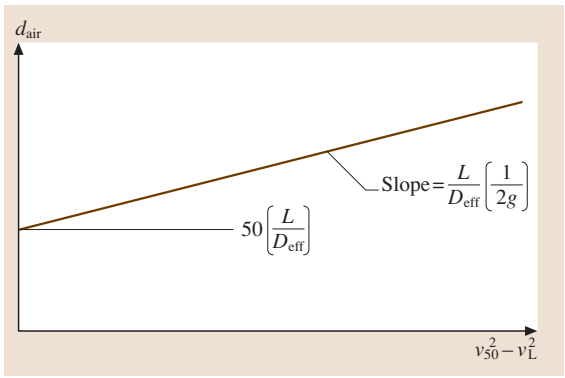


Fig. 13.196 Landing air run determination

speed and is usually about $1.25V_S$. The glide distance is

$$d_{GL} = 50\left(\frac{L}{D_{eff}}\right),$$

where $D_{eff} = D - T$,

$$d_{decel} = \frac{V_{50}^2}{2a} - \frac{V_L^2}{2a} = \frac{\frac{1}{2}\left(\frac{W}{g}\right)V_{50}^2 - \frac{1}{2}\left(\frac{W}{g}\right)V_L^2}{D_{eff}}.$$

Since lift is essentially equal to the weight,

$$d_{air} = 50\frac{L}{D_{eff}} + \frac{1}{2g}\left(V_{50}^2 - V_L^2\right)\frac{L}{D_{eff}}$$
$$= \frac{L}{D_{eff}}\left[50 + \frac{1}{2g}\left(V_{50}^2 - V_L^2\right)\right].$$

L/D_{eff} is the effective L/D ratio during the air run. It can be determined from flight test air runs by plotting flight test air run distances versus $(V_{50}^2 - V_L^2)$, as illustrated in Fig. 13.196.

The ground deceleration distance is

$$d_G = \frac{V_L^2}{2a} = \frac{V_L^2}{2[R/(W/g)]},$$

where

R = effective average resistance or total stopping force
$= \mu(W - L) + D$,
μ = braking coefficient of friction (Table 13.26),
D = drag, including drag of flaps, slats, landing gear, and spoilers.

Note that both d_{air}, air distance, and d_G, ground stopping distance, are directly proportional to V_{50}^2 and/or V_L^2. Both V_{50}^2 and V_L^2 are fixed percentages above V_S for safety reasons. Thus landing distance is linear in V_S^2 except for the glide distance from 50 ft, which depends only on the L/D in the landing configuration.

Table 13.26 Runway friction coefficients

Surface	μ typical values	
	Rolling (brakes off)	Brakes on
Dry concrete/asphalt	0.03–0.05	0.3–0.5
Wet concrete/asphalt	0.05	0.15–0.3
Icy concrete/asphalt	0.02	0.06–0.10
Hard turf	0.05	0.4
Firm dirt	0.04	0.3
Soft turf	0.07	0.2
Wet grass	0.08	0.2

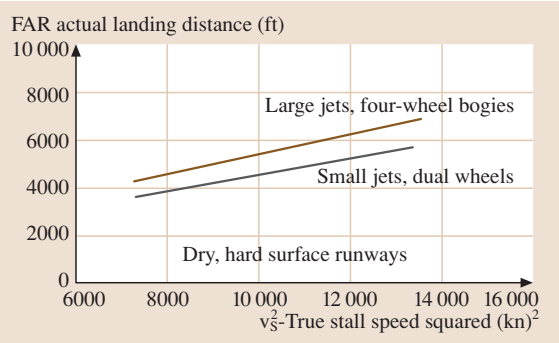


Fig. 13.197 FAR 25 demonstrated landing distance

Thus, for similar airplanes with similar L/D values and equivalent braking systems (i. e., similar μ), landing distances can be of the form

$$d_{\text{land}} = A + BV_S^2 ,$$

where $A = 50(L/D)_{\text{eff}}$.

The FAR 121 scheduled landing field length is defined as the actual FAR 25 demonstrated distance from a 50 ft height to a full stop on a dry, hard surface runway, increased by a factor of 1/0.60, i.e., a 67% increase. Curves of landing field length versus V_S^2 based on flight tests (Fig. 13.197) are linear, but vary between airplanes due to the effective L/D in the air run, the effective coefficient of friction, and the drag in the ground deceleration. It should be noted that the FAR dry runway tests do not allow the use of engine thrust reversers. FAR dry runway distances must be increased by 15% for operation on wet runways.

13.4.10 Stability and Control

The subject of airplane stability and control deals with the ability of an airplane to fly straight with wings level without pilot input (stability) and the ability of the pilot to produce moments about the various airplane axes (control). *Static stability* refers to the initial tendency of the airplane to return or move away from its equilibrium position following a disturbance. *Dynamic stability* is concerned with the entire history of the airplane motion, in particular whether the motion subsides or diverges. In the following discussion, the airplane will usually be treated as a rigid body, a reasonable assumption for the majority of conditions studied. However, at high dynamic pressures many of the structural elements deflect under aerodynamic loads, which further complicates stability and control analysis. The study of the behavior of the structural elements under load and the interaction with aerodynamic load is called *aeroelasticity*. Aeroelastic effects usually tend to reduce static stability. In extreme cases, this interaction may lead to dangerous undamped structural oscillations called *jitter*. The rigid-body motions of an airplane may be divided into two classifications: longitudinal and lateral motions. *Longitudinal motions* occur in the plane of symmetry, whereas *Lateral motions* displace the plane of symmetry. For normal symmetric airplanes, with small displacements from equilibrium, these two types of motions are independent of each other.

Static Longitudinal Stability

Static aerodynamic stability in pitch, more commonly known as static longitudinal stability, is most easily achieved through the use of an aft-mounted horizontal tail. In concept, static longitudinal stability may be defined as the tendency of the airplane to return to its original flight condition without pilot input, when disturbed from steady, unaccelerated flight. While not an absolute requirement for sustained, controlled flight, static longitudinal stability has been found to be a desir-

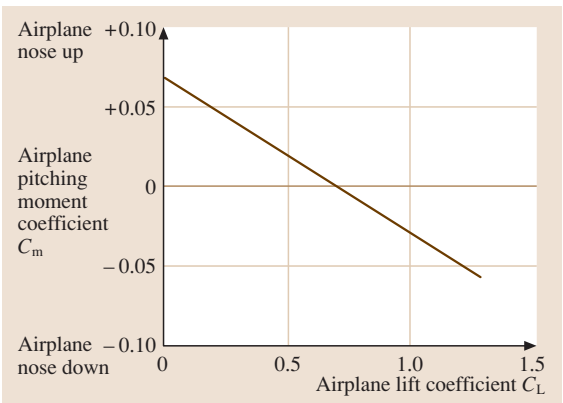


Fig. 13.198 Airplane pitching moment curve

able characteristic for ease of flight, that the variation of airplane pitching moment coefficient C_m with airplane lift coefficient C_L has a negative slope, as shown on the pitching moment diagram in Fig. 13.198.

It should be noted that the pitching moment coefficient is defined with respect to the airplane center of gravity (c.g.) location, expressed as a percentage of the wing mean aerodynamic chord (m.a.c.) aft of the leading edge. For example, with a c.g. location expressed as 25% m.a.c., the c.g. is located at a distance of 25% of the m.a.c. length aft of the leading edge of the m.a.c. Using the sign convention noted earlier, the negative slope of the pitching moment curve means that, for an airplane in equilibrium (pitching moment equal to zero) in steady, un accelerated flight at a given lift coefficient, any disturbance which increases the airplane lift coefficient will result in a negative, or airplane nose down (AND), pitching moment coefficient.

Effect of C.G. Location. It can be shown by analysis of the equation for the airplane pitching moment coefficient versus lift coefficient that the airplane c.g. location has a powerful effect on static longitudinal stability. The results of this analysis show that for every 1% of m.a.c. that the c.g. is moved forward, the slope of the pitching moment curve, dC_m/dC_L about that c.g. will become more negative by 0.01. This point is illustrated for a typical airplane configuration in Fig. 13.199. In this figure, the pitching moment curve with the c.g. located at 25% m.a.c. shows a negative (stable) slope of −0.05. If the c.g. is moved forward to 20% m.a.c., the pitching moment curve about this c.g. location shows a negative (stable) slope of −0.10, indicating that a 5% forward movement in c.g. location results in a pitching moment curve that is more negative by 0.05. Similarly, if the c.g. is moved aft, the pitching moment curve will have a less stable (negative) slope about that c.g. location. In Fig. 13.199, if the c.g. location is moved aft to 30% m.a.c., then the slope of the pitching moment curve about this c.g. location is zero.

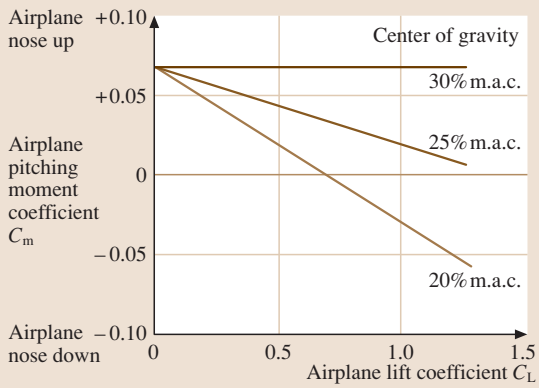


Fig. 13.199 Effect of c.g. location on airplane pitching moments

This situation illustrates another important concept in the discussion of static longitudinal stability, that of aerodynamic center.

Aerodynamic Center. From basic aerodynamics, the aerodynamic center location for any configuration is defined as the center of constant pitching moments, that is, the c.g. location where the pitching moment coefficient remains constant as the lift coefficient varies. In the illustration of Fig. 13.199, the aerodynamic center for the configuration shown is at 30% m.a.c., since for this c.g. location the pitching moment coefficient, C_m, is constant as the lift coefficient varies. For any aircraft configuration, the criterion for achieving static longitudinal stability is that the aircraft c.g. must be located forward of the airplane aerodynamic center (a.c.). This criterion applies to all types of configurations, conventional aft-tail arrangements as well as canards, three-surface layouts, and even flying wings. However, as we shall see, this stability criterion is most easily met using a conventional aft-mounted horizontal tail. Again referring to Fig. 13.199, with the c.g. located at 25% m.a.c. and the a.c. located at 30% m.a.c., this configuration meets the criterion for static longitudinal stability, and with dC_m/dC_L equal to −0.05 is said to be stable by 5% m.a.c. With the c.g. located at 20% m.a.c. and the a.c. at 30% m.a.c., dC_m/dC_L is −0.10 and the configuration is said to be stable by 10% m.a.c. Expressed in equation form

$$dC_m/dC_L = X_{c.g.} - X_{a.c.},$$

where X is expressed in terms of a percentage of the m.a.c.

Aerodynamic Center Buildup. The aerodynamic center location for a complete airplane configuration is determined by the contribution of the various elements of the configuration in pitch. The contributions of these elements can be calculated with reasonable accuracy, but are usually verified by wind-tunnel model tests very early in the preliminary design phase. The major contributors to the complete configuration aerodynamic center are the wing, fuselage, engine nacelles, and the horizontal tail. A typical wind-tunnel model buildup to determine the contributions of these elements to the

complete configuration a.c. is shown in Fig. 13.200. These data are all referred to a c.g. location of 25% m.a.c. By measuring the slope of the pitching moment curve about this c.g., one can obtain the a.c. location for any configuration made up of the major elements that contribute to the a.c. location as

$$X_{\text{a.c.}} = 0.25 - (\mathrm{d}C_\mathrm{m}/\mathrm{d}C_\mathrm{L})\,.$$

Figure 13.200 summarizes the a.c. locations for the various partial configurations, leading up to the complete configuration a.c. Also shown is the contribution of the various elements to the complete airplane a.c.

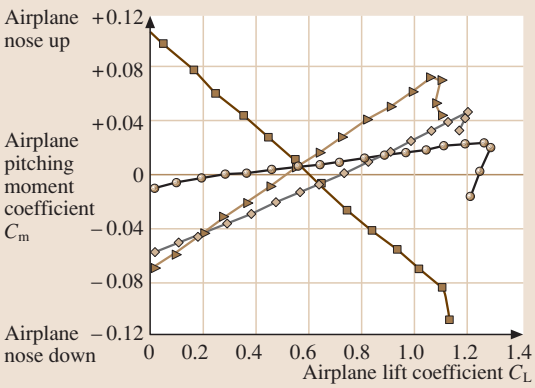
The wing alone a.c. is at 22.6% m.a.c. The wing plus fuselage a.c. is at 16.3% m.a.c., indicating that the fuselage has a destabilizing or unstable contribution of 0.063 or 6.3% m.a.c. The wing plus fuselage plus engine nacelles have an a.c. location of 11.0% m.a.c., indicating that the nacelles have an unstable contribution of 5.3% m.a.c. to the a.c. location. The addition of the horizontal and vertical tails to the model results in an a.c. location for the complete configuration of 42.7% m.a.c. Since the vertical tail has no aerodynamic contribution in pitch, the horizontal tail provides a strong stabilizing contribution of 31.7% m.a.c. Some generalizations may be made from the data. First, the wing-only a.c. is usually around 25% m.a.c., not surprising since the a.c. for nearly all airfoil sections which make up the wing is within 1% m.a.c. or so of the 25% m.a.c. point. Wing sweep may also move the wing alone a.c. a percent point or two, usually aft. Secondly, fuselages are destabilizing contributors, tending to move the a.c. location forward. The larger the fuselage is relative to the wing, the more destabilizing will be its contribution to the complete airplane a.c. Forward-mounted nacelles, like the ones shown on the model in Fig. 13.200, are also destabilizing, although aft fuselage-mounted nacelles are usually stabilizing. Finally, the aft horizontal tail is a major stabilizing contributor to the complete airplane a.c. location. It can be shown that the horizontal tail contribution to static longitudinal stability is dependent on the distance between the 25% chord point on the wing m.a.c. and the 25% m.a.c. point on the horizontal tail, called the horizontal tail length l_H, and the horizontal tail area S_H. This is quite logical since static longitudinal stability involves the generation of aerodynamic restoring moments which are dependent on an aerodynamic force from the horizontal tail (proportional to the horizontal tail area) and a moment arm (proportional to the horizontal tail length).

Longitudinal Control

In addition to providing a major contribution to static longitudinal stability, the horizontal tail is also a source of longitudinal control moments. These control moments are used by the pilot to achieve equilibrium in pitch ($C_\mathrm{m} = 0$) at any desired lift coefficient, allowing control of airspeed in unaccelerated flight, and the curvature of the flight path in accelerated flight. A typical pitching moment diagram illustrating the control moment required to balance the pitching moment at another lift coefficient is shown in Fig. 13.201.

This airplane is stable about its c.g. with a $\mathrm{d}C_\mathrm{m}/\mathrm{d}C_\mathrm{L}$ equal to -0.10, and is in equilibrium ($C_\mathrm{m} = 0$) at a lift coefficient $C_\mathrm{L} = 0.5$. This means that the airplane will fly steadily at a speed corresponding to this lift coefficient, and its static longitudinal stability will resist any disturbances tending to deviate from this speed. If the pilot desires to slow the airplane

Straight wing, four-engine, propeller-driven transport model, center of gravity at 25% m.a.c.



Symbol	Configuration	Notation
—○—	Wing alone	W
—◇—	Wing + Fuselage	W F
—▷—	Wing + Fuselage + Nacelles	W F N
—▭—	Wing + Fuselage + Nacelles + H.T. + V.T.	W F N H V

Configuration	p.m. curve slope $\mathrm{d}C_\mathrm{m}/\mathrm{d}C_\mathrm{L}$	a.c.location % m.a.c.	Effect of element on a.c. location
W	+0.024	22.6	–
W F	+0.087	16.3	$(-6.3\%)_{\text{fuselage}}$
W F N	+0.140	11.0	$(-5.3\%)_{\text{nacelles}}$
W F N H V	−0.177	42.7	$(+31.7\%)_{\text{horiz. tail}}$

Fig. 13.200 Aerodynamic center buildup

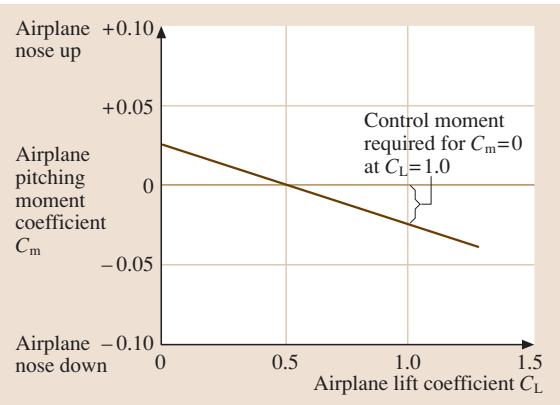


Fig. 13.201 Airplane pitching moment diagram, control moment requirement

down, and fly at a $C_L = 1.0$, he must be equipped with some type of control that can overcome the airplane nose-down moment coefficient of -0.05 at a $C_L = 1.0$, as shown in Fig. 13.201, in order to establish equilibrium at $C_L = 1.0$. Obviously, the more stable the airplane, the more control power that must be provided to change the equilibrium lift coefficient. Thus the designer must achieve a proper balance between the amount of static longitudinal stability provided and the amount of control power available. Longitudinal control power is usually provided through the hinged, moveable aft portion of the horizontal tail (elevators), although in some designs the control power is provided by moving the entire horizontal tail about a fixed pivot point (all-moveable horizontal tail or stabilator).

Longitudinal control capability for a specific configuration may be shown on a pitching moment diagram

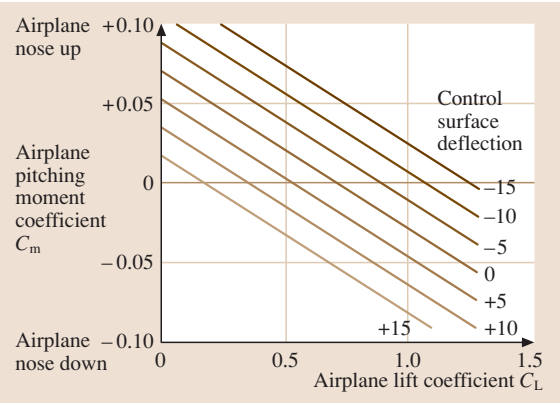


Fig. 13.202 Pitching moment diagram – effect of control deflections

(Fig. 13.202), in which the pitching moment curves with various control surface deflections show the control deflection required to obtain equilibrium at various lift coefficients. Notice that a negative (trailing-edge-up) control deflection produces a positive (airplane-nose-up) pitching moment.

Static Directional Stability

Static directional stability for an airplane is defined as its tendency to develop restoring moments when disturbed from its equilibrium sideslip angle, normally zero. The static directional stability of an airplane is assessed from a chart of yawing moment coefficient, C_n, versus sideslip angle β as shown in Fig. 13.203. Using the sign convention of Figs. 13.155 and 13.157, a positive value of $dC_n/d\beta$ is required for static directional stability; i. e., a positive (airplane-nose-left) sideslip produces a positive (airplane-nose-right) yawing moment.

Similar to the longitudinal case, the static directional stability of an airplane may be determined by adding up the contributions of the various elements of the configuration in sideslip. Analysis has shown that the main contributors to the airplane static directional stability are the fuselage and the vertical tail. The wing, a major element in longitudinal stability, has a negligible effect on the directional stability. This is due to the fact that an angle of sideslip produces very small cross-wind forces on the wing, whereas an angle of attack can produce very large lift forces. The fuselage is a major contributor to static directional stability, and its contribution is always unstable (destabilizing). The stabilizing contributor to static directional stability is the vertical tail, in reality a low-aspect-ratio wing attached to the aft fuselage. The contribution of the major ele-

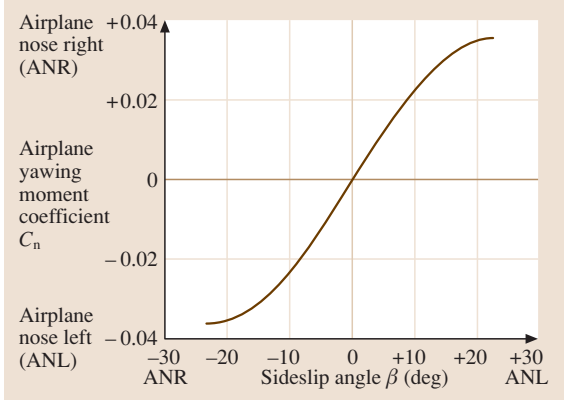


Fig. 13.203 Typical directional stability diagram

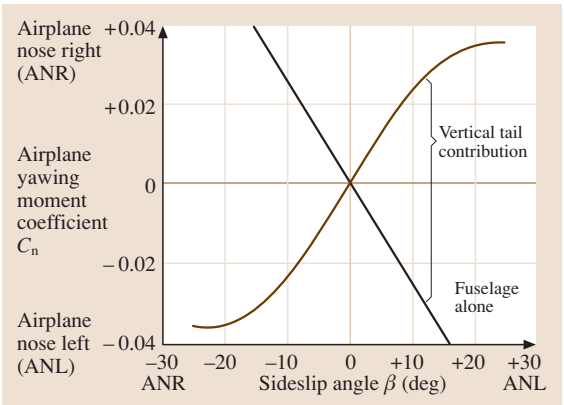


Fig. 13.204 Directional stability buildup

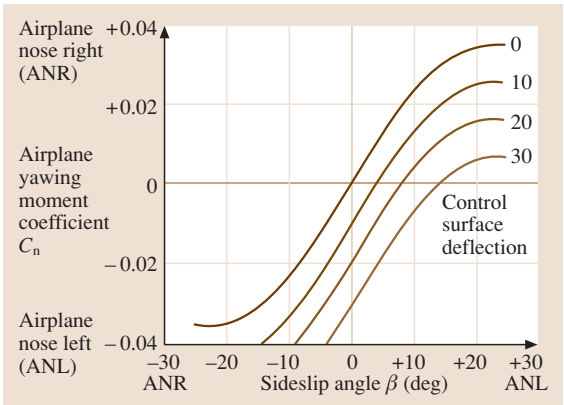


Fig. 13.205 Directional stability diagram – effect of control deflections

ments is shown in Fig. 13.204. When a sideslip angle develops due to a disturbance, the vertical tail experiences an increase in its angle of attack, and produces a restoring moment. The loss in directional stability at high sideslip angles, the *roundover* of the directional stability curve, is due to the vertical tail reaching its maximum lift capability, and stalling as the sideslip angle is increased. This roundover is not desirable and is often offset by the addition of a dorsal fin located at the intersection of the vertical tail leading edge and the fuselage. The magnitude of the restoring moment generated by the vertical tail depends on the distance from the airplane c.g. to the 25% m.a.c. point on the vertical tail, called the vertical tail length l_v and the vertical tail area S_v. For convenience l_v is usually taken from the 25% m.a.c. point on the wing, rather than the c.g.

Directional Control

The vertical tail also provides the means for directional control. The predominant means is through a hinged, moveable aft portion of the vertical tail (rudder), although some advanced military aircraft use all-moveable verticals. As shown in Fig. 13.205, directional control is used to obtain equilibrium ($C_n = 0$) in steady sideslips. This figure shows that 30° of left rudder produces about 14° of positive sideslip.

Another requirement for directional control is to offset the asymmetric thrust moment which develops on a multi-engine airplane when one engine becomes inoperative. In this situation, the aerodynamic moment from the vertical tail at or near-zero sideslip with the control surface deflected must balance the thrust moment caused by the loss of one engine. This situation is most critical at low speeds during take-off. The minimum speed for which it is possible to maintain directional equilibrium ($C_n = 0$) during take-off with one engine inoperative is called the minimum control speed V_{mc}, which may be obtained graphically by the intersection of the yawing moment due to the inoperative engine and the yawing moment due to full directional control, as shown in Fig. 13.206.

A summary of typical pilot control systems for airplanes is shown in Table 13.27.

Longitudinal Dynamics

In order to understand the requirements for static stability and control, it is necessary to study the dynamic

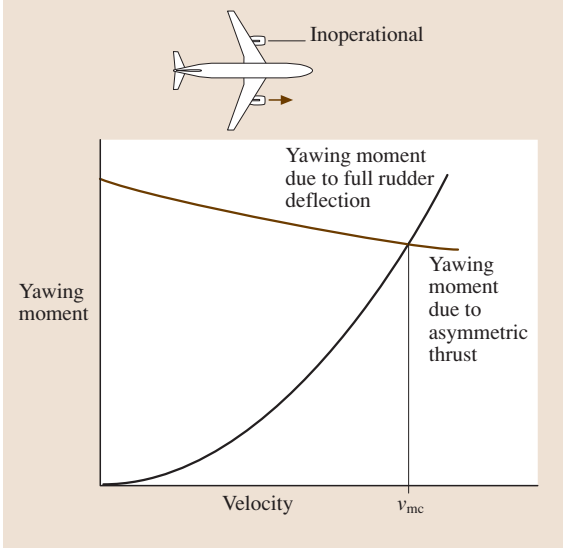


Fig. 13.206 Determination of ground minimum control speed

Table 13.27 Typical pilot control systems

Control about	Cockpit controller	Control direction	Airplane response
Lateral axis	Stick pull Stick push	Pitch	Nose up Nose down
Longitudinal axis	Stick right Stick left	Roll	Bank rwd Bank lwd
Vertical axis	Pedal right Pedal left	Yaw	Nose right Nose left

characteristics of the airplane, investigating the types of motion that characterize the response of the airplane to a disturbance from some equilibrium flight condition and the nature of the transient motions of the airplane to the movement of its controls. Dynamic systems in general have four different modes of motion when responding to a disturbance from an equilibrium condition. These modes are oscillatory or periodic, damped or undamped, as shown in Fig. 13.207.

The characteristic modes for nearly all airplanes are two oscillations, one of long period with poor damping, called the *phugoid*, and one of short period with heavy damping, called the *short period*. The *phugoid* oscillation is one in which there is a large-amplitude variation in airspeed, pitch angle, and altitude The short period is a heavily damped oscillation in which the angle of attack varies at nearly constant speed.

Lateral Dynamics

There are two types of lateral dynamic motions. The first is called the *spiral mode*, which involves variations in bank angle and sideslip. This mode is usually a pure divergence, starting with a slow spiral in the direction of the disturbance, which if uncorrected, will develop into a high-speed spiral dive. The second motion is called the *Dutch roll*, because of its similarity to the well-known ice skating figure. This is an oscillatory motion involving variations in roll and yaw angles that for straight-wing propeller-driven airplanes is usually damped. However for swept-wing airplanes, the Dutch roll oscillation is often lightly damped or mildly divergent, requiring the installation of a supplemental *yaw damper* system to provide satisfactory damping.

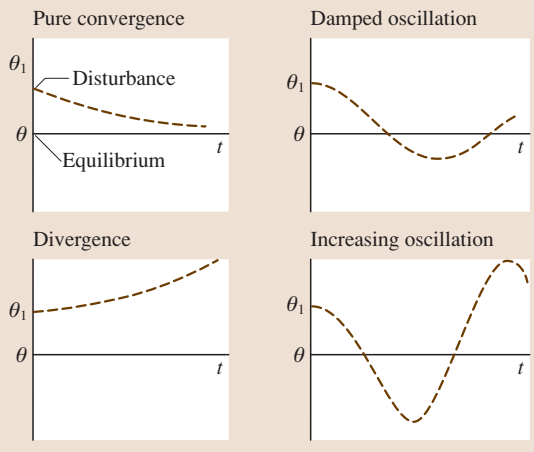


Fig. 13.207 Typical modes of motion

Maneuverability and Turning

The airplane flight path is controlled by varying the magnitude of the lift vector and by varying the output of the power plant. The magnitude of the lift vector is directly related to the lift coefficient through the angle of attack. The pilot can vary the angle of attack by controlling the pitching moment contributed by the control surfaces so that the pitching moment for the complete airplane is zero at the desired lift coefficient. From a steady unaccelerated level-flight condition, a maneuver in the plane of symmetry is initiated by a pilot control input to produce a positive nose-up pitching moment. This nose-up pitching moment will be balanced at some higher lift coefficient by the airplane's static

Load factors in level flight turns

ϕ (deg)	$n = L/W = 1/\cos(\phi)$
0	1.000
15	1.035
30	1.155
45	1.414
60	2.000
75	5.464
90	∞

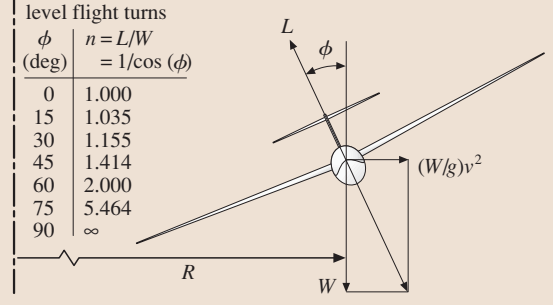


Fig. 13.208 Forces and load factors in level flight turns

longitudinal stability. If the longitudinal control input is made fairly rapidly, the speed will not have time to reduce to the speed required for equilibrium, and the lift will exceed the weight. With lift greater than weight, the airplane will experience a vertical acceleration. The maneuver just described is called an *abrupt pull-up* and may be the start of more complicated maneuvers. For a level-flight turn, referring to Fig. 13.208, the horizontal component of the lift vector accelerates the airplane laterally and curves the flight path. In a turn of radius R, the lateral force $L\sin\phi$, where ϕ is the angle of bank, must balance the centrifugal force on the airplane. Thus

$$L\sin\phi = \frac{(W/g)V^2}{R} \,. \tag{13.5}$$

For a level-flight turn, the weight W must be equal to the vertical component of lift $L\cos\phi$. Substituting in (13.5), we obtain

$$L\sin\phi = \frac{[(L\cos\phi)/g]V^2}{R} \,,$$

$$\tan\phi = \frac{V^2}{gR} \,. \tag{13.6}$$

Equation (13.6) specifies the angle of bank required for any speed and radius of turn. Conversely, the radius of turn is given by

$$R = \frac{V^2}{g\tan\phi} \,.$$

Also, since for a level-flight turn

$$W = L\cos\phi \,,$$

it follows that the lift for such a turn must be given by $L = W/\cos\phi$ and

$$\frac{L}{W} = \frac{1}{\cos\phi} = n \,.$$

As we shall see later, the quantity $n = L/W$ is an important parameter defined as the *load factor*.

13.4.11 Loads

Aircraft structures must be designed to withstand the most serious of the infinite number of possible combinations of external loads that may act on it in flight and when landing. Experience, accumulated over many years of design, analysis, and research, has led to the formulation of a very rational set of procedures that determine the design loads and define the airspeeds for which the design loads are imposed. For civil aircraft, these requirements and procedures are described in the federal air regulations (FAR) part 23 and part 25 *Airworthiness Standards, Airplanes*. For military aircraft, these requirements and procedures are described in MIL-A-8660, *Airplane Strength and Rigidity, General Specification For* and MIL-A-8661, *Airplane Strength and Rigidity, Flight Loads*. The requirements are, in most cases, nearly identical in both the civil and military documents. The information that follows will be based on FAR 23 and 25, with information from MIL-A-8660 and MIL-A-8661 added where significant differences exist:

- Flight conditions (FAR 25.331–25.459):
  - Maneuver load generated by intentional pilot application of controls
  - Gust load generated by sudden change in angle of attack due to encountering a *gust*
- Landing conditions (FAR 25.473–25.511):
  - Level landing
  - Tail-down landing
  - One-wheel landing
  - Side load conditions
  - Braked roll conditions
  - Yawing conditions

Air Loads

Flight Load Factor. An important concept in the analysis of air loads imposed under various flight conditions is the flight load factor n, which is defined as follows

$$n = \frac{\text{aerodynamic force} \perp \text{longitudinal axis}}{\text{aircraft weight}} \,.$$

For an aircraft in steady, level flight, the aerodynamic force perpendicular to the longitudinal axis is given by the lift, which is equal to the weight. Since the weight is due to the force of gravity, the aircraft is said to be in 1 g flight. If the lift is four times the weight, the aircraft is subjected to 4 g's. In a simpler form,

$$n = \frac{\text{lift}}{\text{weight}} \,.$$

V–n Diagrams. The analysis of the critical design air loads for an aircraft employs a chart known as the V–n diagram. These charts show flight load factors as a function of equivalent airspeed and represent the maximum load factors expected in service, based on the requirements of the applicable specifications. These load factors are called *limit* load factors. The airplane structure must withstand these loads without damage.

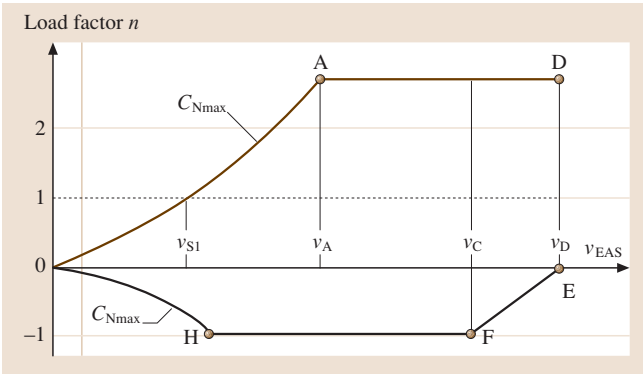


Fig. 13.209 *V–n* diagram: maneuvering envelope

These limit loads are multiplied by a safety factor of 1.5 to define *ultimate* or failure loads. There are two types of *V–n* diagrams; one to define maneuver load factors, and one to define gust load factors.

V–n Diagram: Maneuver Envelope FAR 23.333 and FAR 25.333. The *V–n* diagram showing the maximum maneuver load factors that must be used for structural design (Fig. 13.209) is an envelope defined by various lines and points which have a specific relationship to the design load factors. A brief explanation of the key portions of the maneuvering envelope is given below.

Line 0–A. This line describes the load factor that results when the aircraft is maneuvered to its maximum normal force coefficient $C_{N_{max}}$ in the clean or cruise configuration. Since this is the maximum normal force that can be generated by the aerodynamic characteristics of the configuration, it is the maximum load factor that can be generated by the pilot. The equation of this line is $n = C_{N_{max}} qS/W$.

Point A. This is the intersection of the pull up to $C_{N_{max}}$ with the maximum positive maneuver load factor specified in the requirements for the particular type of aircraft being designed. It should be noted that point A is not selected by the designer, but is determined uniquely by the aircraft parameters and the maneuver limit load factor for the type.

Line A–D. This is the maximum positive maneuver load factor for the type. The design limit load factors for various aircraft types were determined many years ago from flight tests of a number of airplanes of various types, each subjected to a number of typical maneuvers. These tests were made with an accelerometer placed at or near the airplane center of gravity, which recorded the imposed accelerations. Experience has indicated that these load factors resulted in highly satisfactory designs.

Table 13.28 Maximum maneuver load factors

Aircraft type	Max positive	Max negative
Normal	3.8	1.52
Acrobatic	6.0	3.00
Commuter	3.8	1.52
Utility	4.4	1.76

Line 0–H. This line describes the load factor generated when the airplane is maneuvered to its maximum negative $C_{N_{max}}$ value. Since wing design is focused on using airfoils that have high values of positive $C_{N_{max}}$, the maximum values of negative $C_{N_{max}}$ are usually about 0.7 times the positive $C_{N_{max}}$ values.

Line H–F. This line describes the maximum negative maneuver load factor, again determined from flight tests as noted above.

The maximum maneuver load factors vary with aircraft type (Table 13.28). For FAR 23 aircraft, the maximum positive and negative maneuver load factors are listed in Table 13.29.

For FAR 25 aircraft, the maximum positive maneuver load factor varies with design gross weight. The maximum value is 3.8 up to a gross weight of 4100 lb. At higher gross weights, the maximum value varies according to the relation up to a gross weight of 50 000 lb where the maximum becomes a constant value of 2.5. The maximum negative maneuver load factor for FAR part 25 aircraft is −1.0. Corresponding maneuver load factors for military aircraft are shown in Table 13.29.

V–n Diagram: Gust Envelope FAR 23.333 and FAR 25.333. In addition to the load factors imposed by intentional maneuvers controlled by the pilot, appreciable increases in effective angle of attack result from entering a *gust*, or current of air having a velocity component normal to the line of flight. The resulting increase in load factor depends primarily on the vertical velocity of the gust, and especially for business jets and jet transports, it may exceed the maximum due to intentional maneuvers. These load factors are summarized on a *V–n* gust envelope diagram (Fig. 13.210).

The load factors produced by gusts vary directly with equivalent airspeed, and are computed using the

Table 13.29 Design maneuver load factors for military airplanes

Aircraft type	Basic flight design weight		All weights	Max. design weight		Max. ordnance weight	
	Max.	Min. at V_H	Min. at V_H	Max.	Min. at V_H	Max.	Min. at V_H
Fighter/attack (subsonic)	8.00	−3.00	−1.00	4.00	−2.00	5.50	−2.00
Fighter/attack (supersonic)	6.50	−3.00	−1.00	4.00	−2.00	5.50	−2.00
Observation trainers	6.00	−3.00	−1.00	3.00	−1.00		
Utility	4.00	−2.00	0	2.50	−1.00		
Tactical bomber	4.00	−2.00	0	2.50	−1.00		
Strategic bomber	3.00	−1.00	0	2.00	0		
Assault transport	3.00	−1.00	0	2.00	0		
Conventional transport	2.50	−1.00	0	2.00	0		

gust load factor equation given in FAR 23.341 and FAR 25.343.

The equation is

$$n = 1 + \frac{K_g U_{gE} V_E a}{498(W/S)},$$

where

$$K_g = \frac{0.88\mu_g}{5.3 + \mu_g} = \text{gust alleviation factor},$$

$$\mu_g = \frac{2(W/S)}{\rho \overline{C}_{ag}} = \text{airplane mass ratio},$$

U_{gE} = equivalent gust velocity (ft/s),

ρ = density of the air (slug/ft^3),

W/S = wing loading (lb/ft^2),

$\overline{C}$ = mean geometric chord (ft),

g = acceleration due to gravity (ft/s^2),

V_E = aircraft equivalent airspeed (kts),

a = slope of the airplane normal force curve per radian.

The V–n diagram for the gust envelope is shown in Fig. 13.210.

The designer must assume a symmetrical vertical gust of:

- 66 fps at V_B from sea level (S.L.) to 20 000 ft, decreasing to 38 fps at 50 000 ft
- 50 fps at V_C from S.L. to 20 000 ft, decreasing to 25 fps at 50 000 ft
- 25 fps at V_D from S.L. to 20 000 ft, decreasing to 12.5 fps at 50 000 ft

The key points of the gust envelope are as follows.

Line 0–B′. As in the maneuver envelope, this line describes the maximum load factor that can be generated by a gust which causes the airplane to reach its $C_{N_{max}}$.

Point B′. This point is the intersection of the load factor for $C_{N_{max}}$ and the load factor for a 66 fps gust. This point determines V_B, the design speed for maximum gust intensity.

Point C′. This point is the intersection of the load factor due to a 50 fps gust and the design cruising speed V_C.

Point D′. This point is the intersection of the load factor due to a 25 fps gust and the design dive speed, V_D.

Points E′, F′, and G′. These points are the corresponding intersections for negative gusts at the designated speeds.

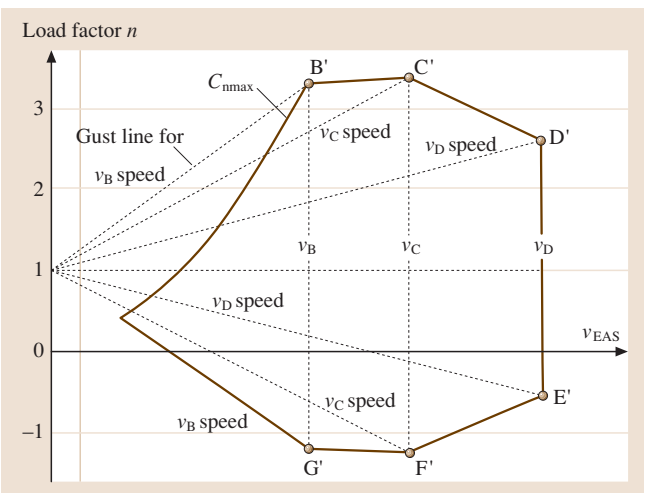


Fig. 13.210 V–n diagram: gust envelope

The maneuver and gust envelopes are superimposed to determine the highest load factors for design at all speeds within the flight envelope and the entire aircraft structure analyzed for these load factors.

In addition to these major air loads imposed on the airplane by intentional maneuvers and gusts, there are other conditions associated with abrupt control inputs in pitch, roll, and yaw that must be accounted for in the structural design.

Design Airspeeds

The airspeeds associated with the V–n diagram except for V_A and V_B are chosen by the designer, but must meet certain definitions and criteria contained in the FARs. The following list is a simplified summary:

Design airspeeds – EAS:

V_S – Stalling speed or minimum steady flight speed;
V_A – Maneuver speed or full control deflection speed;
V_B – Design speed for maximum gust intensity;
V_{FE} – Design flap-extended speed;
V_{LE} – Design landing-gear-extended speed;
V_{LO} – Design landing-gear-operating speed (if different from V_{LE});
V_C – Design cruising speed ($\geq V_B + 43$ kts);
V_{MO} – Maximum operating limit speed (*Barber Pole* speed);
V_{FC} – Maximum speed at which flight characteristics requirements must be met;
V_D – Design dive speed, $\geq V_C/0.80$, or speed reached in 7.5° dive for 20 s from V_C, followed by $1.5g$ recovery.

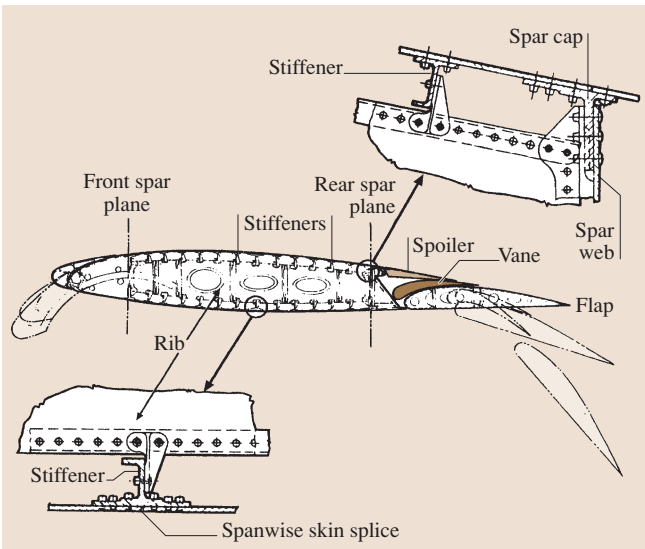


Fig. 13.211 Wing box structural elements

The military speed definitions are basically the same, although MIL-A-8660B combines V_C and V_{MO} into a maximum level flight speed V_H, and replaces V_D with the *limit speed* V_L.

For subsonic airplanes, the design airspeeds are usually constant for the entire flight envelope. For high-subsonic and supersonic airplanes, the design airspeeds are varied throughout the flight envelope, since equivalent airspeeds that are appropriate at sea level and lower altitudes are beyond the performance capabilities of the airplane at higher altitudes. Therefore, the design airspeeds for these types are usually defined in terms of Mach number at higher altitudes; for example, the maximum operating limit speed is defined by a V_{MO}/Ma_{MO} line that is a function of altitude. As noted earlier, the design cruise speed (Mach number) need not be higher than the maximum speed in level flight at that altitude with maximum cruise power. This provision usually sets Ma_C. V_{MO} is usually set at or slightly above (0.01 or 0.02 Mach number) Ma_C, providing a margin on the order of 0.08 Mach number between the best long-range cruise Mach number and Ma_{MO}. The design dive Mach number is usually about 0.05 Mach number higher than Ma_{MO}.

Ground Loads

There are a range of take-off and landing conditions that must be considered in the structural design of the airplane. These conditions are described in detail in the appropriate sections of the federal aviation regulations. These loads may be classified as vertical loads due to descent rates at touchdown and taxiing over rough surfaces, longitudinal loads caused by wheel spin-up loads on landing, braking loads, and rolling friction loads, and lateral loads caused by landing with some sideslip angle, cross-wind taxiing, and turning on the ground.

13.4.12 Airplane Structure

The structure of an aircraft must withstand the applied aerodynamic and ground reaction loads encountered in normal operation, as well as those that may be encountered very rarely. The essential character of aircraft structure is light weight, because weight plays such an important role in the performance and economics of the

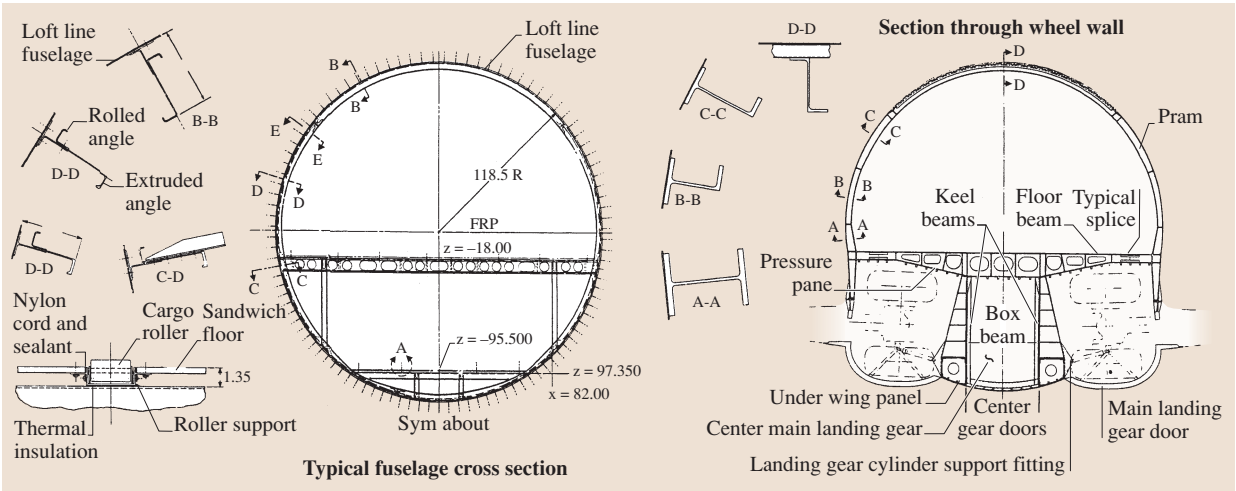


Fig. 13.212 Fuselage structural elements

airplane. This critical significance of structural weight is the major difference between aircraft structural design and other types of structural design.

Aircraft structural design has always sought to meet the applied load requirements with a minimum acceptable margin of safety and the least weight. However, the potentially disastrous effect of an aircraft structural failure requires that the structure must be designed for long life either with safe life criteria or fail-safe design. Safe life means that the stresses in a component are so low that fatigue failure is not possible over the anticipated life of the airplane, or at least until some

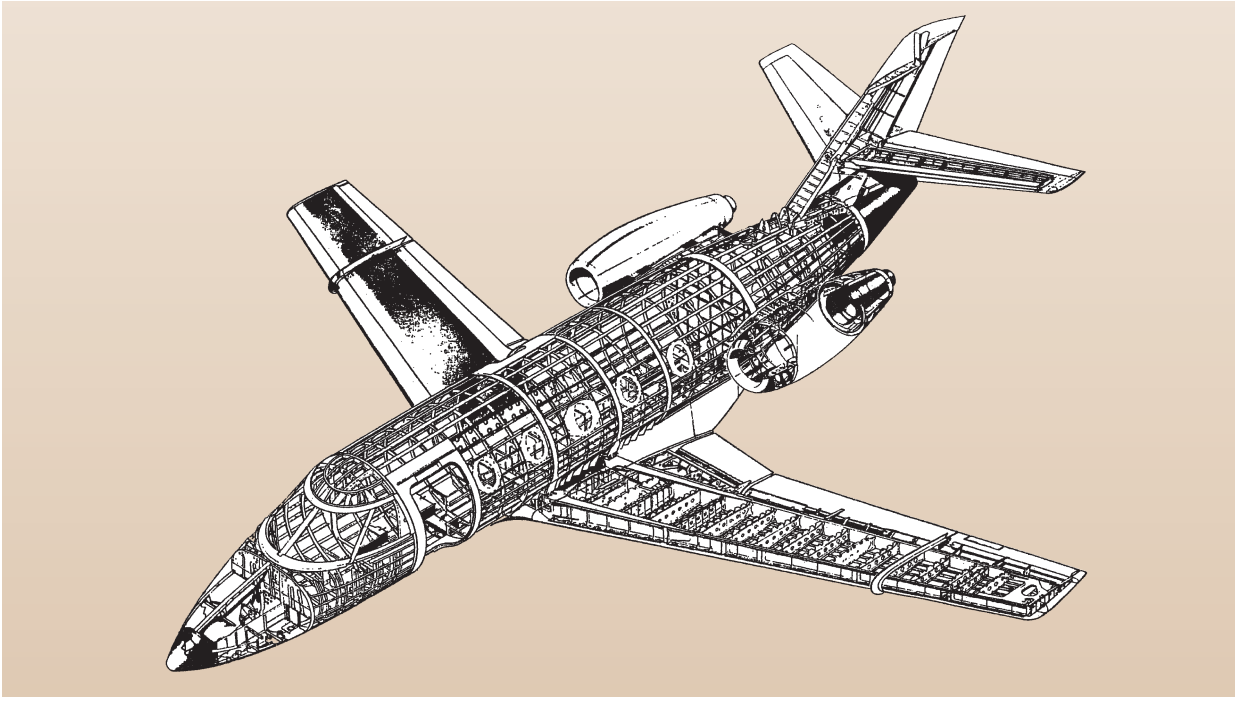


Fig. 13.213 Complete airplane structure overview

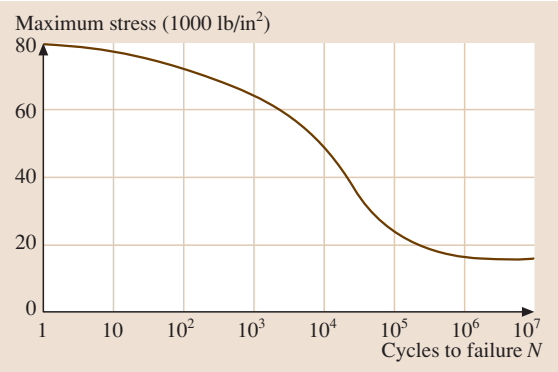


Fig. 13.214 Fatigue test results

period has passed after which a part replacement is required. Fail-safe means that the structure has alternate load paths so that no single failure of a part will be hazardous to the airplane. This is accomplished by designing the structure so that no one part carries most of the maximum load. Therefore, if one part fails, the remainder of the structure can still carry most of the maximum load. Since the maximum load is rarely encountered, and the structure has a safety factor of 1.5, the structure remains safe until the failure is found and repaired.

Structural Design

Aircraft structures are composed of three basic types of structural elements: stiffened shells, stiffened plates, and beams. The terms stiffened shells and stiffened plates refer to the fact that, under compressive loading, the thin skins of the shells and plates generally buckle before reaching the compressive failure stress of the material. To avoid this condition, stiffeners are attached to the thin skins. Stiffeners not only carry their own load, but by preventing early buckling of the skin, they increase the stress that can be supported in compression before buckling occurs. The primary wing structural element, the wing *box*, is essentially a hollow beam, consisting of the upper and lower wing skins, ribs, wing skin stiffeners, spar caps, and spar webs, as shown in Fig. 13.211.

The wing skins transfer the external air loads to the wing box structure. The fuselage structure (Fig. 13.212) consists of frames which maintain the cross-sectional shape, to which are attached various longitudinal stiffeners called stringers or longerons. Loads on the fuselage floor are supported by floor beams, with other beam elements to accommodate cutouts in the fuselage shell. Horizontal and vertical tail structure is very similar to the wing structure, employing skins, ribs, stiffeners, spars, spar caps, and spar webs.

An overview of the complete structure for a large transport airplane is shown in Fig. 13.213.

Another important aspect of structural design is the need to design for long fatigue life. Metals suffer a gradual deterioration under repeated application and removal of loads. Modern commercial transports may fly for many thousands of hours and take-off and landing cycles, so that often fatigue life, rather than strength requirements, dominate the structural design. The number of load cycles a material can tolerate depends on the stress level. The lower the stress level, the greater number of cycles the part can withstand. The results of a typical fatigue test (Fig. 13.214) show the tremendous improvement in fatigue life that can be obtained by limiting a material to cyclic stress levels that are well below the ultimate strength.

If fatigue is critical, a less strong material with better fatigue resistance may result in a lighter structure. Excessive stress levels conducive to early fatigue failure may arise not only from an overall high stress level, but also from stress concentrations at local points in the structure. Fittings and joints, which serve to carry loads from one structural component to another, may, if not designed very carefully, introduce increased local stresses leading to fatigue failure long before any problem occurs in the basic structure. A major part of the design of aircraft structure is the avoidance of stress concentrations by careful detail design. One approach is to lower the stress level approaching a hole or fitting

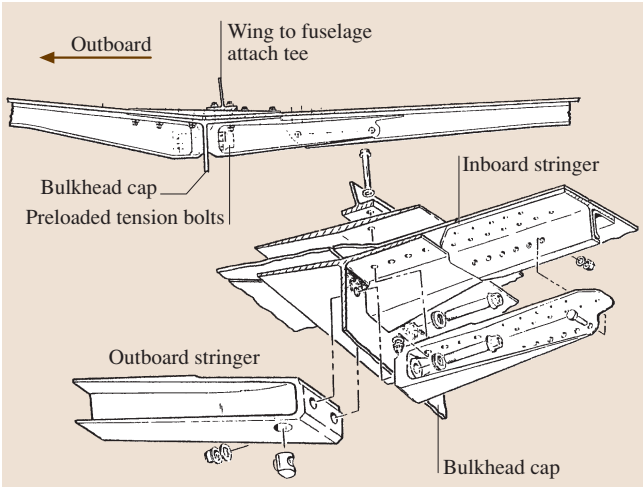


Fig. 13.215 Wing–fuselage joint fitting

by adding an extra sheet or thickness of metal called a *doubler* to the inside surface of a wing skin around an access hole or to the fuselage skin around a window. Special attention is given to fitting design (Fig. 13.215). In spite of the care taken in detailed structural design, a vital part of aircraft maintenance is the continuing search on a scheduled basis for structural cracks.

Structural Analysis

In the design and analysis of real aircraft structure, usually a large assemblage composed of various structural elements such as stiffened shells, stiffened plates, and beams, the overall geometry becomes extremely complex, and cannot be represented by a single mathematical expression. In addition, these built-up structures are characterized as having material and structural discontinuities such as cutouts, thickness variations in the members, as well as discontinuities in loading and support structure. It is apparent that classical methods can no longer be used, particularly those which require the formulation and solution of governing differential equations. For complex structures, the preferred method of analysis is called the finite element stiffness method. With the advent of high-speed large-storage-capacity digital computers, finite element matrix methods have become the most widely used tools in the analysis of complex structures.

Structural Materials

Modern aircraft are constructed from a variety of materials, chosen based on considerations such as density, mechanical properties, corrosion resistance, ease of fabrication, and cost. The most used materials are the light metals, aluminum alloys, and titanium, although for some applications where high strength is required, steel alloys are used. Structural composite materials are being used more because of their low density and good mechanical properties. Composites generally consist of a plastic matrix of epoxy resin, reinforced by many high-strength fibers of carbon, Kevlar, glass, or boron. The distribution of materials used in aircraft construction has been changing over the years as material science researchers have developed more attractive products. Figure 13.216 shows a comparison of the structural materials distribution between the Boeing 747 of 1969 and the Boeing 777 of 1994. In the 25 years between the design of these two transport aircraft, the most notable difference in materials use is the large

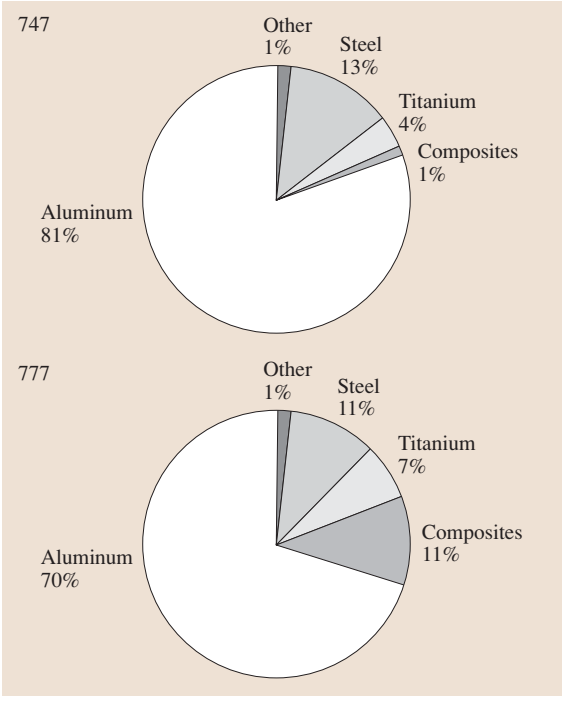


Fig. 13.216 Comparison of structural materials distribution – commercial transports

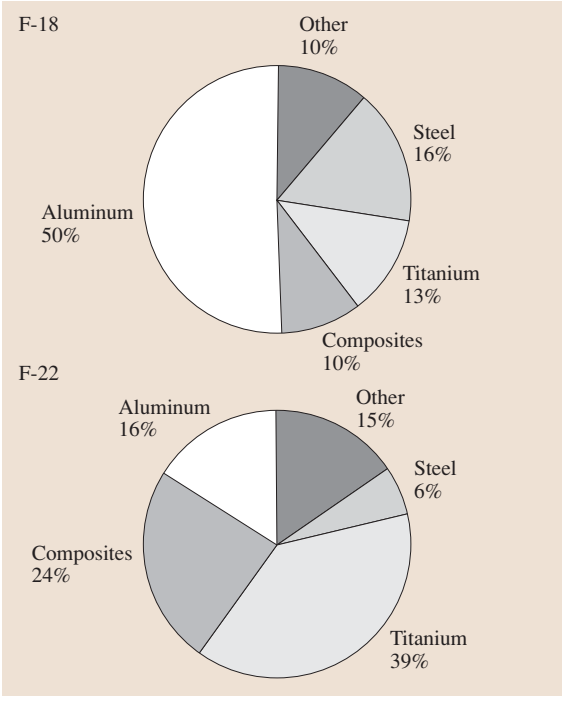


Fig. 13.217 Comparison of structural materials distribution – military fighters

Table 13.30 Summary of typical maintenance checks for transport airplanes

	B 737	B 747	A 300	A 320	Time required	Man-hour
A-Check	350 h	650 h	350 h	350 h	Overnight	20–130
B-Check	5.5 months	1800 h	1000 h	–	One day	200–1000
C-Check	15 months	18 months	18 months	15 months	A few days	600–1400
D-Check	22000 h	31000 h	25000 h	–	≈ 6 weeks	50000
	25000 loading	72 months	12500 loading	–		
	108 months	–	108 months	102 months		

increase in composites and titanium, and a significant reduction in the use of aluminum.

An even more startling change in structural materials distribution has taken place in military aircraft design. Figure 13.217 shows the distribution of structural materials in the McDonnell Douglas F-18 of 1978 and the Lockheed Martin F-22 of 2003. The most startling difference in the materials use between these two aircraft is the large reduction in the use of aluminum alloy and steel in the F-22 and the large increase in the use of titanium and composites.

In addition to the introduction of newer, nontraditional materials, several new processes for producing parts from these new materials have been introduced, which reduce the number of parts required and also reduce the amount of labor required to produce each part. Examples of these new processes are *resin transfer molding* (RTM) for producing composite parts, and *hot isostatic pressing* (HIP) for producing large high-quality castings.

13.4.13 Airplane Maintenance Checks

As noted earlier, in order to insure safe operations, especially for commercial transport aircraft, regular formal maintenance checks are made of the airplanes systems and structure. These checks range from the very simple *preflight* check of the aircraft conducted by the flight crew prior to each flight, to the extensive, detailed maintenance inspection checks conducted by technical specialists at major maintenance facilities. A summary of the various types of maintenance checks that are conducted by the airline operators is presented in Table 13.30.

References

13.1 J.Y. Wong: *Theory of Ground Vehicles* (Society of Automotive Engineers, Warrendale 1993)
13.2 H. Heisler: *Advanced Vehicle Technology* (Butterworth-Heinemann, Oxford 1989)
13.3 D. Hoyle: *Automotive Quality Systems Handbook* (Butterworth-Heinemann, Oxford 2000)
13.4 C. Clarke: *Automotive Production Systems and Standardization* (Springer, Berlin Heidelberg 2005)
13.5 D. Gruden (Ed.): *Traffic and Environment* (Springer, Berlin Heidelberg 2003)
13.6 IEA: *Energy Balance of OECD Countries 2005* (IEA, Paris 2005)
13.7 IEA: *Energy Balance of Non-OECD Countries 2005* (IEA, Paris 2005)
13.8 Energy Information Administration (EIA): *International Energy Annual 2004* (EIA, Washington 2006), http://www.eia.doe.gov
13.9 IFEU: *UmweltMobilCheck, Wissenschaftlicher Grundlagenbericht* (IFEU – Institut für Energie und Umweltforschung, Heidelberg 2006), in German
13.10 Lufthansa: *Balance – Das Wichtigste zum Thema Nachhaltigkeit bei Lufthansa* (Lufthansa, Frankfurt 2007), in German
13.11 E. Pasanen: *Ajonopeudet ja jalankulkijan turvallisuus [Driving Speeds and Pedestrian Safety]* (Helsinki University of Technology, Espoo 1991), in Finish
13.12 K. Eichner: *Auto Jahresbericht 2004* (Verband der Automobilindustrie e.V. (VDA), Frankfurt 2004), http://www.vda.de/de/service/jahresbericht/files/VDA2004.pdf, in German
13.13 D. Clausing: *Total Quality Development* (ASME, New York 1994)
13.14 D. Bastow, G. Howard, J.P. Whitehead: *Car Suspension and Handling* (Society of Automotive Engineers, Warrendale 2004)
13.15 H.-H. Braess, U. Seiffert (Eds.): *Vieweg Handbuch der Kraftfahrzeugtechnik* (Vieweg, Wiesbaden 2003), in German
13.16 http://www.nhtsa.gov
13.17 J. Happian-Smith (Ed.): *An Introduction to Modern Vehicle Design* (Butterworth-Heinemann, Oxford 2002)
13.18 L.L. Beranek: *Noise and Vibration Control* (McGraw-Hill, New York 1971)

13.19 A.D. Dimaragonas, S. Haddad: *Vibration for Engineers* (Prentice Hall, Englewood-Cliffs 1992)
13.20 B. Hall: Noise vibration and harshness. In: *An Introduction to Modern Vehicle Design*, ed. by J. Happian-Smith (Butterworth-Heinemann, Oxford 2002)
13.21 T.D. Gillespie: *Fundamentals of Vehicle Dynamics* (Society of Automotive Engineers, Warrendale 1992)
13.22 J. Reimpell, H. Stoll, J.W. Betzler: *The Automotive Chassis* (Butterworth-Heinemann, Oxford 2001)
13.23 W.F. Milliken, D.L. Milliken: *Race Car Vehicle Dynamics* (Society of Automotive Engineers, Warrendale 1995)
13.24 W.F. Milliken, D.L. Milliken: *Chassis Design: Principles and Analysis* (Society of Automotive Engineers, Warrendale 2002)
13.25 H.B. Pacejka: *Tyre and Vehicle Dynamics* (Butterworth-Heinemann, Oxford 2002)
13.26 C. Campbell: *Automobile Suspensions* (Chapman Hall, London 1981)
13.27 J.R. Ellis: *Vehicle Handling Dynamics* (Mechanical Engineering Ltd., London 1994)
13.28 J.C. Dixon: *Tires, Suspension and Handling* (Society of Automotive Engineers, Warrendale 1996)
13.29 ISO: *ISO 8855: Road Vehicles – Vehicle Dynamics and Road-Holding Ability – Vocabulary* (International Organization for Standardization, Geneva 1991)
13.30 J.D. Halderman, D.C. Mitchell: *Automotive Steering, Suspension, and Alignment* (Prentice Hall, Englewood-Cliffs 2003)
13.31 P.C. Brooks, D.C. Barton: Braking Systems. In: *An Introduction to Modern Vehicle Design*, ed. by J. Happian-Smith (Butterworth-Heinemann, Oxford 2002)
13.32 D. Neudeck, R. Martin, N. Renzow: Porsche brake development: from the race track to the road. In: *Advanced Brake Technology*, ed. by B.J. Breuer (Society of Automotive Engineers, Warrendale 2003)
13.33 T.P. Newcomb, R.T. Spurr: *Braking of Road Vehicles* (Chapmann & Hall, London 1996)
13.34 R. Limpert: *Brake Design and Safety* (Society of Automotive Engineers, Warrendale 1992)
13.35 W.C. Orthwein: *Clutches and Brakes: Design and Selection* (Marcel Dekker, New York 2004)
13.36 U. Stoll: SBC – The Electro-Hydraulic Brake System from Mercedes-Benz. In: *Advanced Brake Technology*, ed. by B. Breuer, U. Dausend (Society of Automotive Engineers, Warrendale 2003)
13.37 J.B. Heywood: *Internal Combustion Engine Fundamentals* (McGraw-Hill, New York 1988)
13.38 F. Schleder: *Stirlingmotoren* (Vogel, Würzburg 2002), in German
13.39 K. Mollenhauer: *Handbuch Dieselmotoren* (Springer, Berlin Heidelberg 1997), in German
13.40 C.R. Ferguson, A.T. Kirkpatrick: *Internal Combustion Engines: Applied Thermosciences* (Wiley, Hoboken 2001)
13.41 C.F. Taylor: *The Internal Combustion Engine in Theory and Practice* (MIT Press, Boston 1985)
13.42 W.W. Pulkrabek: *Engineering Fundamentals of the Internal Combustion Engine* (Prentice Hall, Englewood-Cliffs 2003)
13.43 R. Stone: *Introduction to Internal Combustion Engines* (Society of Automotive Engineers, Warrendale 1999)
13.44 H. Heisler: *Advanced Engine Technology* (Society of Automotive Engineers, Warrendale 1995)
13.45 F. Zhao, D.L. Harrington, M.-C. Lai: *Automotive Gasoline Direct-Injection Engines* (Society of Automotive Engineers, Warrendale 2002)
13.46 L. Guzzella, C.H. Onder: *Introduction to Modeling and Control of Internal Combustion Engine Systems* (Springer, Berlin Heidelberg 2004)
13.47 SAE: Advanced Hybrid Vehicle Powertrain Technology, SAE 2002 World Congress, Detroit, Michigan (Society of Automotive Engineers, Warrendale 2002)
13.48 G. Killmann: Toyota Prius – Development and market experiences, VDI Bericht 1459 (VDI, Düsseldorf 1999)
13.49 P. Eastwood: *Critical Topics in Exhaust Gas Aftertreatment* (Taylor & Francis, London 2001)
13.50 G. Mom: *The Electrical Vehicle: Technology and Expectations in the Automobile Age* (Johns Hopkins Univ. Press, Baltimore 2004)
13.51 R. Johansson, A. Rantzer (Eds.): *Nonlinear and Hybrid Systems in Automotive Control* (Springer, Berlin Heidelberg 2003)
13.52 R.L. Evans: *Automotive Engine Alternatives* (Springer, Berlin Heidelberg 1987)
13.53 J.T. Pukrushpan, A.G. Stefanopoulou, H. Peng: *Control of Fuel Cell Power Systems* (Springer, Berlin Heidelberg 2004)
13.54 N.D. Vaughan, D. Simner: *Automotive Transmissions and Drivelines* (Butterworth-Heinemann, Oxford 2002)
13.55 P.G. Gott: *Changing Gears: The Development of the Automotive Transmission* (Society of Automotive Engineers, Warrendale 1991)
13.56 G. Lechner, H. Naunheimer: *Automotive Transmissions: Fundamentals, Selection, Design, and Application* (Springer, Berlin Heidelberg 1999)
13.57 T. Birch, C. Rockwood: *Automatic Transmissions and Transaxles* (Prentice Hall, Englewood-Cliffs 2001)
13.58 T. Birch: *Automotive Heating and Air Conditioning* (Prentice Hall, Englewood-Cliffs 2002)
13.59 W. Fung, M. Hardcastle: *Textiles in Automotive Engineering* (CRC, Boca Raton 2001)
13.60 H. Wallentowitz, C. Amsel (Eds.): *42 V-PowerNets* (Springer, Berlin Heidelberg 2003)
13.61 U. Seiffert, L. Wech: *Automotive Safety Handbook* (Society of Automotive Engineers, Warrendale 2003)
13.62 E. Chowanietz: *Automobile Electronics* (Newnes, Burlington 1995)

13.63 R.K. Jurgen (Ed.): *Automotive Electronics Handbook* (McGraw-Hill, New York 1995)
13.64 C.O. Nwagboso (Ed.): *Automotive Sensory Systems* (Chapman Hall, Boca Raton 1993)
13.65 U. Kiencke, L. Nielsen: *Automotive Control Systems: For Engine, Driveline and Vehicle* (Springer, Berlin Heidelberg 2000)
13.66 J.F. Keshaw, J.D. Halderman: *Automotive Electrical and Electronic Systems* (Prentice Hall, Englewood-Cliffs 2004)
13.67 J. Marek, H.-P. Trah, Y. Suzuki, I. Yokomori (Eds.): *Sensors for Automotive Technology* (Wiley, Hoboken 2003)
13.68 J. Valldorf, W. Gessner (Eds.): *Advanced Microsystems for Automotive Applications 2005* (Springer, Berlin Heidelberg 2005)
13.69 T. Rybak, M. Steffka: *Automotive Electromagnetic Compatibility (EMC)* (Springer, Berlin Heidelberg 2004)
13.70 B. Peacock, W. Karwowski (Eds.): *Automotive Ergonomics* (Taylor & Francis, London 1993)
13.71 J.M. Porter, C.S. Porter: Occupant accommodation: an ergonomics approach. In: *An Introduction to Modern Vehicle Design*, ed. by J. Happian-Smith (Butterworth-Heinemann, Oxford 2002)
13.72 M. Blomè, T. Dukic, L. Hanson, D. Högberg: Simulation of Human-Vehicle Interaction in Vehicle Design at Saab Automobile: Present and Future. In: *Recent Developments in Automotive Safety Technology*, ed. by D. Holt (Society of Automotive Engineers, Warrendale 2004) pp. 621–627
13.73 M. Huang: *Vehicle Crash Mechanics* (CRC, Boca Raton 2002)
13.74 N. Jones: *Structural Impact* (Cambridge Univ. Press, Cambridge 1997)
13.75 J.A.C. Ambrosio, M.F.O.S. Pereira, F.P. da Silva (Eds.): *Crashworthiness of Transportation Systems: Structural Impact and Occupant Protection* (Springer, Berlin Heidelberg 1997)
13.76 F.J. Stützler, C. Chou, J. Le, P. Chen: Development of CAE-Based Crash Sensing Algorithm and System Calibration. In: *Recent Developments in Automotive Safety Technology*, ed. by D. Holt (Society of Automotive Engineers, Warrendale 2004) pp. 327–337
13.77 G. Pahl, W. Beitz: *Engineering Design* (Springer, Berlin Heidelberg 1997)
13.78 L. Sage: *Winning the Innovation Race: Lessons from the Automotive Industry's Best Companies* (Wiley, Hoboken 2001)
13.79 M. Maurer, C. Stiller (Eds.): *Fahrerassistenzsysteme mit maschineller Wahrnehmung* (Springer, Berlin Heidelberg 2005), in German
13.80 G. Döllner, C. Gümbel, O. Tegel: Prozesse und Bausteine des CAx-Datenmanagements in der Digitalen Produktentwicklung, 6. Automobiltechnische Konferenz 'Virtual Product Creation 2002' (Berlin 2002), in German
13.81 C.E. Armi: *American Car Design Now: Inside the Studios of America's Top Car Designers* (Rizzoli, New York 2004)
13.82 A. Parnow: Nutzen und Einsatz von RAMSIS bei DaimlerChrysler, RAMSIS User Conference (Kaiserslautern 2004), in German
13.83 M.B. Abbott, D.R. Basco: *Computational Fluid Dynamics* (Longman, Harlow 1989)
13.84 R.H. Barnard: *Road Vehicle Aerodynamic Design* (MechAero, St. Albans 2001)
13.85 W.H. Hucho (Ed.): *Aerodynamics of Road Vehicles – from Fluid Mechanics to Vehicle Engineering* (Society of Automotive Engineers, Warrendale 1998)
13.86 U. Weidmann: *Grundlagen zur Berechnung der Fahrgastwechselzeiten*, Vol. 106 (IVT, Zürich 1995), in German
13.87 K. Endmann: Bewährung des Y-Stahlschwellenoberbaus, EI–Eisenbahningenieur **10**, 25–30 (2000), in German
13.88 A. van Wilcken, F. Fleischer, H. Lieschke: Herstellung Feste Fahrbahn Rheda, Type Walter-Heilit with bibloc sleeper used for Köln-Rhein/Main, with 300 km/h regular train speed, Eisenbahntechn. Rundsch. **51**, 172–182 (2002), in German
13.89 C. Esveld: *Modern Railway Track*, 2nd edn. (MRT-Productions, Zaltbommel 2001)
13.90 K. Popp, W. Schiehlen: *Fahrzeugdynamik* (Teubner, Stuttgart 1993), in German
13.91 Arbeitsgemeinschaft Rheine-Freren: *Rad/Schiene-Versuchs- und Demonstrationsfahrzeug Definitionsphase R/S-VD* (Ergebnisbericht der Arbeitsgruppe Lauftechnik, Minden 1980), in German
13.92 M. Hecht: New freight bogie is an important contribution for growth of rail-freight, Eur. Railw. Rev. **4**, 61–64 (2002)
13.93 K.H. Grothe, J. Feldhusen, H. Dubbel: *Taschenbuch für den Maschinenbau*, 21st edn. (Springer, Berlin Heidelberg 2005) pp. Q50–Q87, in German
13.94 M. Löber, S. Schneider, N. Sifri, P. Trosch: Innovative crashfähige Kastenstruktur der TRAXX-Lokomotiven, Elektr. Bahn. **102**(H8/9), 334–344 (2004), in German
13.95 J. Pachl: *Systemtechnik des Schienenverkehrs, Bahnbetrieb planen, steuern und sichern*, 4th edn. (Teubner, Stuttgart 2004), in German
13.96 P. Argüelles, J. Lumsden, M. Bischoff: *European Aeronautics: A Vision for 2020* (Office for Official Publications of the European Communities, Luxembourg 2001)
13.97 D.P. Raymer: *Aircraft Design: A Conceptual Approach* (American Institute of Aeronautics and Astronautics, Reston 2006)
13.98 D. Küchemann: *The Aerodynamic Design of Aircraft* (Pergamon, Oxford 1978)
13.99 Greener by Design: http://www.greenerbydesign.org.uk/home/index.php, accessed 18 Oct 2007

13.100 C.B. Millikan: *Aerodynamics of the Airplane* (Wiley, New York 1941)
13.101 A.M. Kuethe, J.D. Schetzer: *Foundations of Aerodynamics* (Wiley, New York 1950)
13.102 H.W. Liepmann, A.E. Puckett: *Aerodynamics of a Compressible Fluid* (Wiley, New York 1947)
13.103 L.M. Nicolai: *Fundamentals of Aircraft Design* (METS, San Jose 1984)
13.104 J.M. Swihart: *Design Choice and Marketing of Commercial Jet Airplanes* (Boeing Commercial Airplane Company, Chicago 1978)
13.105 R.D. Schaufele: *The Elements of Aircraft Preliminary Design* (Aries Publications, Santa Ana 2000)
13.106 Federal Aviation Administration: *Code of Federal Regulations, Title 14, Aeronautics and Space* (Office of the Federal Register: Federal Aviation Administration, Washington 1997)
13.107 L.K. Loftin Jr: *Subsonic Aircraft: Evolution and the Matching of Size to Performance* (NASA Reference Publication 1060, Arlington 1980)
13.108 D.P. Raymer: *Aircraft Design: A Conceptual Approach* (AIAA, Washington 1989)
13.109 J. Roskam: *Airplane Design: Part I, Preliminary Sizing of Airplanes* (Roskam Aviation and Engineering, Ottawa 1989)
13.110 I.H. Abbott, A.E. Von Doenhoff: *Theory of Wing Sections* (Dover, New York 1959)
13.111 R.D. Schaufele, A.W. Ebeling: *Aerodynamic Design of the DC-9 Wiing and High Lift System, SAE Paper No. 67-0846* (Aeronautics and Space Engineering Meeting, Los Angeles 1967)
13.112 Anonymous: *The DC-9 Handbook* (Douglas Aircraft Co., Long Beach 1991)
13.113 Anonymous: *The DC-10 Handbook* (Douglas Aircraft Co., Long Beach 1986)
13.114 C.D. Perkins, R.E. Hage: *Airplane Performance Stability and Control* (Wiley, New York 1949)
13.115 N.S. Currey: *Aircraft Landing Gear Design: Principles and Practices* (AIAA, Washington 1988)
13.116 J. Roskam: *Airplane Design, Part IV; Layout Design of the Landing Gear and Systems* (Roskam Aviation and Engineering Corporation, Ottawa 1989)
13.117 R.S. Shevell, I. Kroo: *Introduction to Aircraft Design Synthesis and Analysis, Course Notes* (Stanford University, Palo Alto 1981)
13.118 Anonymous: *The Aircraft Gas Turbine Engine and its Operation* (United Technologies Corporation, East Hartford 1988)
13.119 Anonymous: *The Jet Engine* (Rolls Royce pic, Derby 1986)
13.120 A.P. Fraas: *Aircraft Power Plants* (McGraw-Hill, New York 1943)
13.121 USAF Stability and Control Datcom: *Air Force Flight Dynamics Laboratory* (Wright-Patterson Air Force Base, Dayton 1975)
13.122 Anonymous: *Brief Methods of Estimating Airplane Performance, Report No. SM-13515* (Douglas Aircraft Co., Santa Monica 1949)
13.123 Anonymous: *DC-9-30 Performance Handbook* (Douglas Aircraft Co., Long Beach 1969)
13.124 H.H. Cherry, A.B. Croshere Jr.: An Approach to the Analytical Design of Aircraft, SAE Quart. Trans. **2**(1), 12–18 (1948)
13.125 K.H. Grote (ed.): *Dubbel Taschenbuch für den Maschinenbau*, 21st edn. (Springer Verlag, Berlin Heidelberg 2002), in German
13.126 D.J. Peery, J.J. Azar: *Aircraft Structures* (McGraw-Hill, New York 1982)
13.127 R.S. Shevell: *Fundamentals of Flight* (Prentice Hall, Englewood Cliffs 1989)

14. Construction Machinery

Eugeniusz Budny, Mirosław Chłosta, Henning Jürgen Meyer, Mirosław J. Skibniewski

In this chapter the most common classes of machinery found on construction sites will be presented. For the purpose of this chapter the authors focus on construction machinery and equipment applications in the building and public utility sectors of the construction industry. The classes of machinery and equipment for earth, concreting, assembly, and finishing works described in this chapter are used not only in these two construction industries, but also in road, bridge, and railway building; pile, tunnel, and water foundation; the opening of mines; the building of natural gas and petroleum pipelines; sewerage systems; cooling towers for the power industry; and other industrial building structures.

One should note that specialized equipment ensuring the efficiency, high quality, and safety of work during the realization of structures is used in almost all these kinds of construction. Even a brief description of this equipment would require a separate publication. For example, in road building alone 63 types of machines (see the draft International Standard ISO/FDIS 22242) are used. In the final part of this chapter the state of automation and robotization of construction machinery is presented.

14.1 Basics

14.1.1 Role of Machines in Construction Work Execution

Construction work is to a large extent hard and labor intensive and often poses a health hazard. This is due to the fact that building production involves handling large masses of materials and that some of the materials, such as lime, paints, industrial chemical products, and asbestos, are detrimental to human health. In addition, most construction works are carried out in the open.

The mechanization of construction started with hard physical work and labor-consuming effort, such as earth works and the horizontal and vertical transport of materials, only later embracing finishing works.

Construction work mechanization is inseparably linked with the technologies used in construction and so one can distinguish work mechanization by the different branches of construction (e.g., housing, public utility building, civil engineering, industrial building, and power facility building) or by particular kinds of work, e.g., earth work, assembly, and finishing work.

The transition from craft methods to a more economically effective form of industrial building, using a wide range of prefabricated units, gave an impetus to the development of more efficient machines for the assembly of such units as well as machinery and equipment for concrete works.

The mechanization of construction is a worldwide phenomenon and determines the development of this industry.

Through construction mechanization one can achieve the following goals:

- Speed up the rate of work in comparison with manual methods and so shorten the construction cycle.
- Reduce labor consumption, increase production capacity, and reduce work costs.
- Make work in construction less arduous and so more attractive.
- Improve work safety (construction is the most hazardous field of human activity).

Particularly rapid advances in construction mechanization were made after World War II in response to the urgent need to increase construction production capacity to provide the population with housing and improve standard of living. This was done by developing the construction machine building industry and new technologies consisting of the assembly of building structures from prefabricated units. Construction mechanization covered the following kinds of work: earth work, vertical and horizontal materials transport, materials handling, assembly, and finishing work (including external and internal plastering, painting, terrazzo and wooden floor sanding, and element fixing).

Mechanization is the primary factor contributing to an increase in productivity and it should be economically worthwhile, which is achieved through the intensive use of modern machines in good working order.

Definitions

Construction machine: a device whose function is to increase or replace human or animal physical force in carrying out construction processes.

In the field of construction machinery one can distinguish two general classes of machines: technological machines (for processing raw materials and/or semifinished products) and transport machines (for changing the location of building elements and materials).

Construction work technology: a method for carrying out construction work.

Mechanization of construction: activities covering the application of machines, equipment, and mechanized tools to carry out manufacturing processes in construction.

14.1.2 Development of Construction Machinery – Historical Outline

The available historical data indicate that cranes form the oldest class of construction machines invented by humans. Cranes were known in Greece as early as the

5th century BC. They were used for transporting structural elements vertically and horizontally and putting them in specified places, among other tasks, in the construction of magnificent temples. Greek cranes incorporated structural components such as toothed and worm gears, pulley blocks, rope drums, supports, and power units in the form of levers mounted directly on the hoisting winch's shaft, or various treadmills.

Figure 14.1 shows a derrick commonly used in both ancient Greece and Rome, with a mast in the form of the letter "A". The hoisted block is grabbed by a crampon. The winch is turned by means of two levers and the mast's inclination is adjusted by guy-ropes.

As late as the second half of the 19th century, i. e., until the industrial revolution, cranes were driven by people or, less commonly, by animals. Treadmills, including treadwheels with steps to be climbed, or pole windlasses were used as the driving devices [14.1] (Fig. 14.2).

Historically, the next class of construction machines to be invented was pile-drivers. These were needed to build structures (including temples) on weak ground and bridges. The first pile-drivers were very similar in their design to cranes. A simple medieval pile-driver consisted of a tripod, a rope sheave, and a rope with a hook on which a heavy stone block was hung. The dropped block would drive a pile into the ground. Devices working on the double-arm lever principle and a hoisting winch were employed to pull out such piles.

Towards the end of the 19th century pile-drivers with a guided drop weight began to be used. Later floating pile-drivers powered by a paddle waterwheel were introduced. Also the pulling up of the pile hammer was improved: the work of several people hoisting the pile hammer to its upper position by simply pulling the rope was replaced by a treadwheel or a treadmill.

The next class of construction machines developed were dredgers [14.2], which were used to build waterways and deepen ports and canals. The first record of the design of a gouge-dipper dredger dates from 1420, presented by Venetian Giovanni Fontana (Fig. 14.3).

The Fontana dredger was installed on a pontoon and its working tool was a dipper moving in a chute with a sharp metal tip. First the lowered dipper was inserted into the ground under its deadweight and then the winning was detached by the gouge-dipper as the latter was pressed into the ground. The winning was brought to the surface by pulling the dipper up in the chute.

The gouge-dipper dredger was the precursor of a whole family of dredgers. Later scraper dredgers,

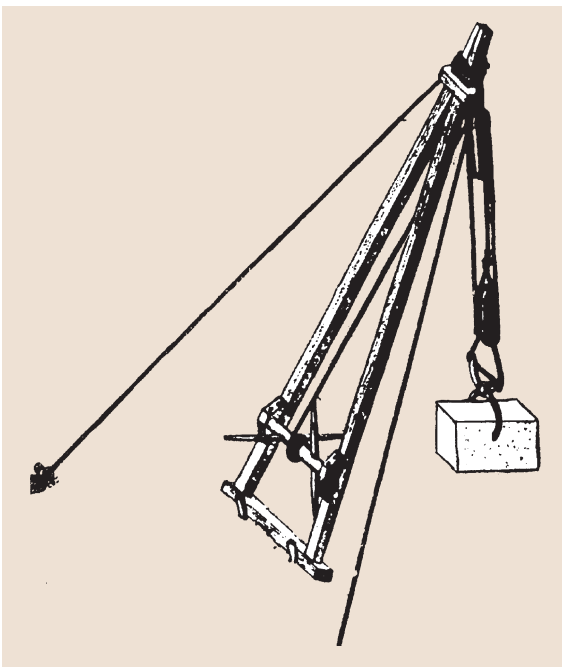


Fig. 14.1 Derrick used in ancient Greece and Rome

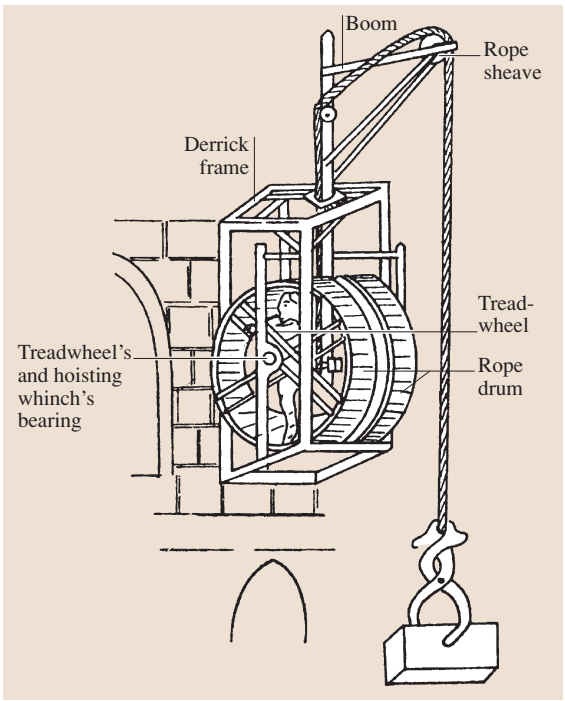


Fig. 14.2 Treadwheel-type boom derrick placed on erected building's wall

clamshell dredgers, dipper dredgers, dipper-scraper dredgers, wheel dredgers, and carding dredgers were designed.

The working mechanisms were driven via hoisting winches by treadwheels, pole windlasses or treadmills powered by men. The exception was a horse-driven 16 m^3/h-capacity dredger with a 0.66 m^3-capacity dipper, built in the USA around 1820.

The development of dipper dredgers gave French designers the idea of putting similar machines to work on land, and so excavators came into being [14.2]. The first excavator designs, although unrealized, appeared in the first half of the 17th century [14.2].

The beginnings of the development of construction mechanization and a breakthrough in construction machine building are associated with the industrial revolution in the UK in the middle of the 18th century.

The most important device in the industrial revolution was the steam engine, improved by Thomas Newcomb (1683–1729) and James Watt (1736–1819). As a result of the Industrial Revolution and its spread to other countries, rapid population growth combined with the expansion of large cities and the development of road, water, and sea transport occurred. The massive and rapid construction of railway lines, roads, canals, and ports involved large-scale earth works. Construction machines, initially powered by steam engines and later combustion engines and electric motors, appeared. The first construction machines in which power drives were employed were dredgers, most commonly bucket-ladder dredgers, used for extracting the winning from the bottom of rivers and canals and for building ports. They were first built in the UK and then in the USA, Russia, and France, and worked in conjunction with loading bridges, transport-tow vessels, and rail transport for carrying off the winning.

The development of the railway system boosted demand for related construction works, track, and rolling stock. Embankments, tunnels, and bridges were built, and tracks were also laid.

In the development of construction machinery one can distinguish the following landmarks:

- The development of the first steam-powered earth-moving machine in 1776. This was a dipper dredger used for dredging the canals in the port of Sunderland.
- The building of the first steam bucket-ladder dredger by Samuel Bentham in 1802 for work in the port of Portsmouth (UK).
- The construction of a floating bucket-ladder dredger, called the *Amphibious Digger*, by Oliver Evans – the pioneer constructor of steam engines in America – for dredging the river and port of Philadelphia in the years 1800–1804.
- The invention of the steel cable by Albert in Germany in 1834, the use of which greatly contributed to the development of cranes.
- The building of the first single-bucket excavator, called the *American Steam Excavator* or the *Yankee Geologist*, by William Smith Otis in the USA in 1836 (Fig. 14.4). Otis's machine has been a symbol of construction mechanization to this day. The proof of its originality and technical excellence is the fact that it set the direction of the development of excavators for 140 years. Otis's first machines were used in the construction of the Baltimore–Ohio railroad.

Otis's excavator with a 1.15 m^3-capacity bucket replaced the work of 80 diggers. Its high economic efficiency significantly contributed to the development of the excavator, grab-dredger, and crane building industry.

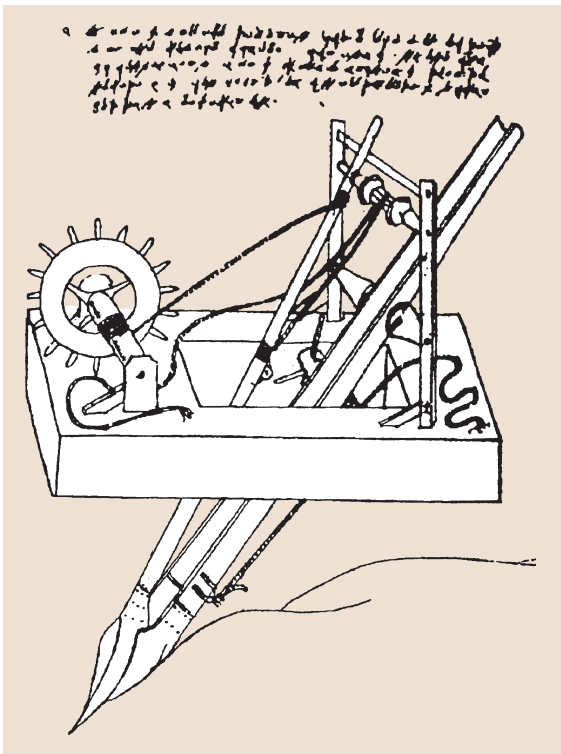


Fig. 14.3 Giovanni Fontana's gouge-dipper dredger

The total production of excavators and dredgers in the USA in 1880 reached 1000 units per annum, and a considerable number of them were exported:

- In 1850 Briton William Fairbairn (UK) developed an arch crane jib made from two riveted plates and then steam-driven cranes.
- In 1861 in Germany Nicolaus August Otto developed a four-stroke, spark-ignition combustion engine.
- In 1874 the company the Aveling & Porter (UK) developed the first steam-driven wheeled crane, called *Little Tom*. In the same year a gantry crane with a truss bridge girder was constructed in north Germany.
- Bucket-ladder dredgers, loading bridges, locomotives, and steam-powered transport-tow vessels were built in France for the construction of the Suez Canal in the years 1865–1869.
- The construction of the Manchester–Liverpool Canal in the UK in the years 1887–1898. Fifty-eight British-made (Ruston, Smith, Whiteker, and Wilson) single-bucket excavators, 18 Priestman clamshell excavators, and many bucket-ladder excavators and dredgers were used to carry out earthwork amounting to 38 million m^3.
- In 1897 Rudolf Diesel built the first compression-ignition engine.
- The development of bucket ladder excavators in France, Germany, and the USA in the second half of the 19th century.
- In 1890 the Osgood Company (USA) built the first electrically driven railway excavator.
- Around 1900 the first traveling tower cranes with a hoisting capacity of 0.25–5.5 Mg were built in Germany and France.
- The building of the Panama Canal under the control of the USA in the years 1904–1914. About 100 railway single-bucket excavators were with a bucket capacity of 2–2.9 m^3 each were used on the building site. They were American-made machines supplied mainly by the Bucyrus Steam Shovel and Dredge Co. (South Milwaukee) and the Osgood Dredge Co. (Troy, NY).

Besides these excavators many other steam-driven machines, such as dredgers and narrow-gauge railways for transporting the output, were used in the construction of the canal.

The 20th century brought rapid development of construction machines. Steam engines were replaced by combustion engines and electric motors, which were

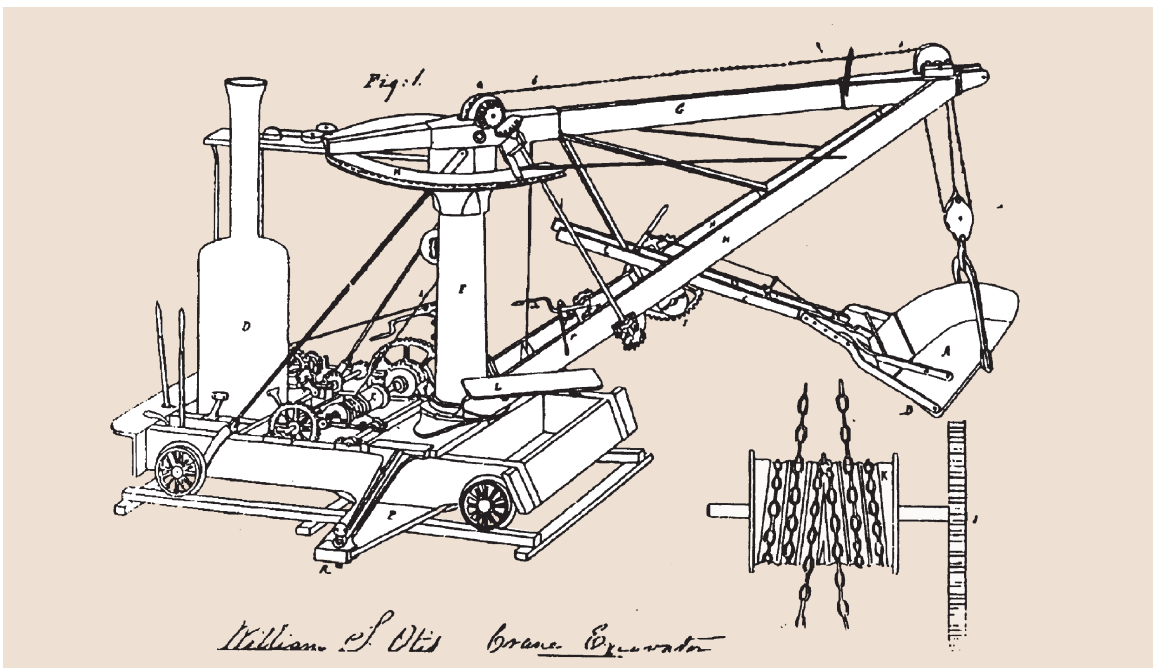


Fig. 14.4 Patent drawing of excavator signed by W. S. Otis on 24 February 1839

easier to operate. High-strength structural materials, hydraulic and pneumatic drives, and control systems incorporating electronics, computer technology, and microprocessors were introduced. As a result of these innovations the performance of construction machines improved and their range of application was extended.

American companies have contributed greatly to the development of construction machinery. For example, Powling and Harnischfeger was one of the first companies to manufacture crawler cranes at the end of World War I and Bucyrus-Erie was the first company in the world to present in 1946 a traveling hydrocrane with a telescopic jib. Several classes of modern construction machines will be described in the next sections of this chapter.

14.1.3 Classification of Construction Machinery

The various construction machines can be divided into two classes. One class comprises specialized machinery exclusively for construction-assembly work while the other includes general-purpose machinery used in various industries. This means that one should distinguish between the terms: *construction machinery* and *machinery used in construction*.

In many cases, however, this distinction is not clear cut. This is so, for example, in the case of motor transport means (trucks, semitrailers, trailers, and tractors), wheeled cranes used for reloading different materials, and power shovels, bulldozers, and front loaders used also in surface mining and in loose-material reloading yards. Another example is the use of helicopters for the assembly of tower structures.

In the literature on this subject one can come across various methods of classifying construction machinery, usually according to the kinds of construction work for which it is used, its function or design.

The classification found in [14.3] can be regarded as the most reliable classification of construction machinery and equipment. This document is not yet an international standard but may become one in the future.

In this report, construction machinery and equipment are divided according to the kinds of work for which they are used into the following seven classes:

1. Earthmoving machinery and equipment
2. Foundation engineering and soil compaction machinery and equipment
3. Machinery and equipment for manufacturing, transporting, and processing concrete and mortar and for reinforcement and formwork
4. Lifting and access machinery and equipment (scaffolds and work platforms)
5. Specialized machinery and equipment used in building construction (e.g., machinery for roadworks and pipe-laying)
6. Equipment for installation, finishing work, and maintenance
7. General-use machinery and equipment used in construction processes

The above classes are divided into subclasses and then into types. Readers interested in the detailed classification are referred to the report itself.

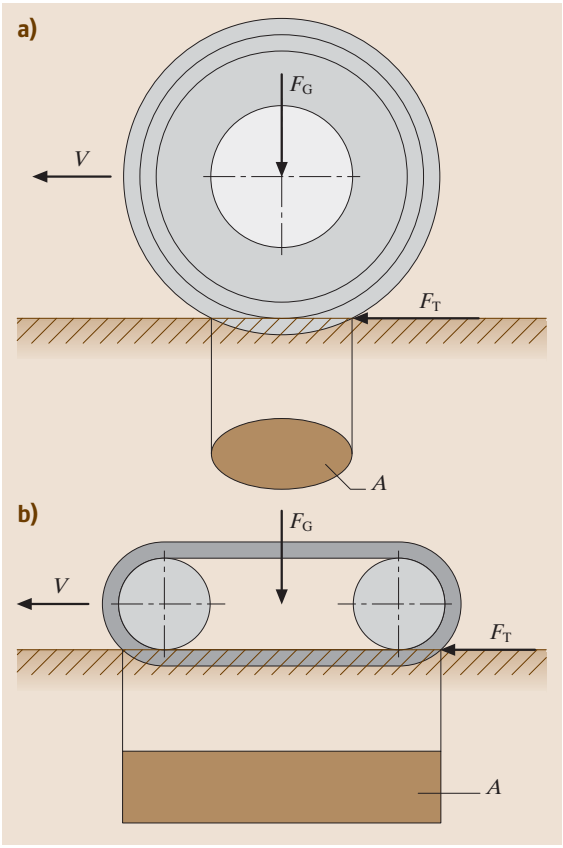


Fig. 14.5a,b Comparison of a wheel **(a)** and a tracklaying chassis' **(b)** footprints

14.2 Earthmoving, Road Construction, and Farming Equipment

It is decisive for human action to use tools, devices, and machines. For a long time man has been making use of machines for construction and agriculture. The huge buildings of antiquity could never have appeared without the application of construction machinery. However, agricultural machinery – especially those according to the EU machinery directive [14.4] – did not emerge until the 18th and 19th century, with the UK and the USA being the first to make use of them [14.5]. A particular problem to be addressed for agricultural machinery was the necessity to be mobile, and today it is still mobility which is one of the most important functions of earthmoving, road construction, and agricultural machines. This is the feature which distinguishes them from stationary machines such as machine tools. Therefore, they are summarized as *mobile working machines*. Due to their requirement for mobility, the following issues are important for this kind of machine and their construction:

- The machine must be connected to an energy source.
- Energy is needed for running the device or the drive.
- The machine can move actively (automotive) or passively (pulled).
- It must be equipped with a (wheel or chain) chassis to supply the tools with energy.
- It must be possible to integrate all components of the machinery system (drive, sensors, actors, etc.) into the vehicle.

Another important feature of earthmoving, construction, and agricultural machinery is the fact that they are designed to fulfil specialized working functions and to be part of complex processes in the area of construction or agriculture.

14.2.1 Soil Science and Driving Mechanics

Soil Composition

Regarding mobile working machines, the soil is of special importance in two regards:

- Mobile working machines move on the soil. Every vehicle requires ground to move on. Depending on the type of chassis used, the drive force must be introduced into the ground in some form, which means that the soil has to take up the driving power; otherwise the wheels or chains would spin. In a broader sense, even roads made of asphalt or concrete must be regarded as soil, representing artificial, manmade hard soils (rocks) that provide a good driving surface.
- Soil is the main resource for plant production and, indirectly, for animal production, thus serving man's demands. Regarding its important role, soil cannot be replaced by any other resource. Soil fertility is mainly dependent on the activity of soil bacteria, fungi, algae, and other microbes. Mobile working machines, especially agricultural machines, are closely linked to this resource, as they are utilized for plant and animal production. The same is true for forestry machines and municipal machines. Earthmoving machines, which are also mobile working machines, also work mainly on soil.

A more detailed description of soil follows, in terms of both its physical and biological features.

Soil consists of various components:

- Solids
- Liquids (water)
- Gas (air)

This distribution of solid, air, and water can be represented as a three-phase system. Soils are classified depending on their aim and application. So, in the area of earthmoving, for rock production, the decisive criteria are different from those in the area of agriculture, which concentrates on soil fertility. An example of agricultural soil classification is the *World Soil Classification* initialized by the Food and Agriculture Organization (FAO) [14.6]. Rock may be classified according to the ISO standard 14689-1 entitled *Geo-technical Investigation and Testing – Identification and Classification of Rock* [14.7]. Apart from these international classification systems there are various national ones [14.8].

Chassis Types

The chassis can be regarded as an interface between the mobile working machine and the ground; two main groups can be distinguished (Fig. 14.5):

- Wheel chassis
- Tracklaying or crawler chassis

In addition to these two groups, there are special-purpose chassis types such as walking chassis to move machines, e.g., in open-cast mining. The driving resistance can be calculated by multiplying the force due to

gravity by the tractive resistance. The most important difference between a wheel chassis and a tracklaying chassis is that a chain tread has a higher driving resistance than a wheel chassis. This has a positive effect on force transmission between the wheel and the soil, reducing soil pressure. However, a wheel chassis permits higher driving speeds than a tracklaying chassis, which turns out to be advantageous, and, with their simpler construction, they can also be used for vehicle suspension.

Wheel Chassis

Force transmission between the wheel and the soil is very complex, being determined by various parameters:

- Penetration into the soil (i. e., soil load-bearing capacity and structure)
- The variety and construction of wheel
- Load and its distribution over area
- Air pressure
- Suspension
- Tyre profile

The roll resistance force value determines the distance between the two vertically acting forces weight and vertical acting force in the case of the driven wheel, and also the angle between the weight and vertical acting force with the pulled wheel (Fig. 14.6).

When the wheel is driven and, additionally, when the traction forces F_T have to be transmitted, e.g., by pulling a trailer or a plough, the propulsion torque is in balance with the soil force.

Figure 14.6 shows the forces acting on a pulled wheel in a simplified way. The driving torque is determined as follows:

$$M = F_T r + F_G f \,, \qquad (14.1)$$

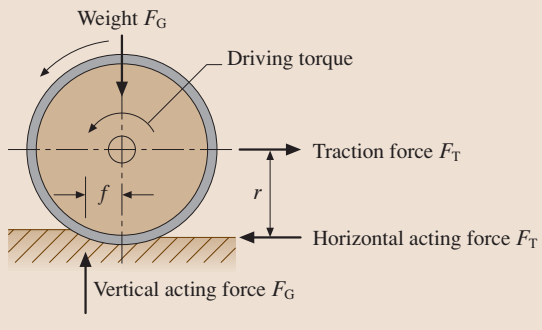


Fig. 14.6 Simplified forces model for a rigid wheel (after [14.9])

where M is the driving torque, F_T is the driving power, F_G is the weight, and r is the effective roll radius. The tractive resistance ρ of a four-wheel tractor is 0.13–0.18 on sand. On concrete or asphalt the tractive resistance is 0.015 [14.9].

Figure 14.7 shows the relation between the wheel perimeter force and the slip. The wheel perimeter force depends on the roll resistance and the driving power.

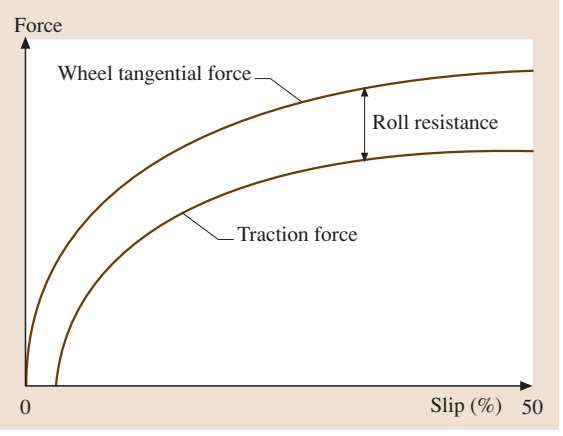


Fig. 14.7 Roll resistance force F_R and wheel traction force F_T depending on the slip σ [14.9]

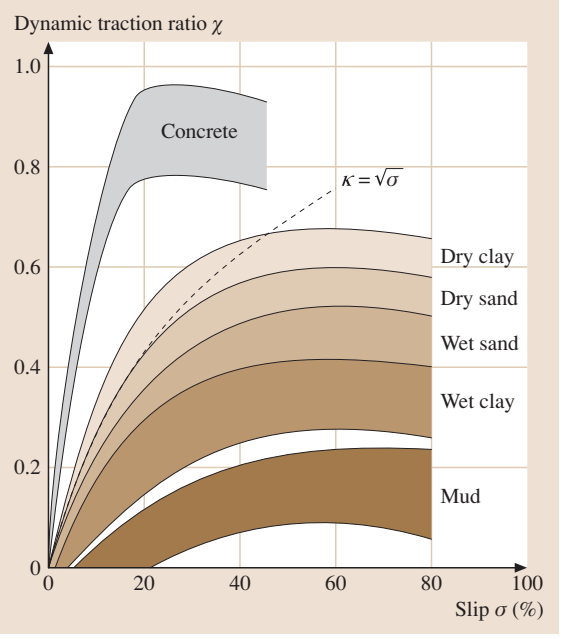


Fig. 14.8 Dynamic traction ration with different soil types (after [14.10])

There is always slip because the roll resistance force F_R has to be overcome to move the vehicle. Indeed, the slip is always small and positive (Fig. 14.7) and, in general, driving is not possible without slip.

The dynamic traction ratio expresses the ratio between the maximum traction force and the wheel load (including the wheel's own weight)

$$\kappa = \frac{F_T}{F_G} . \qquad (14.2)$$

The dynamic traction ratio is determined by measurement. Above all, it depends on the slip σ, as shown in Fig. 14.8. As is shown by this curve, the softer the soil, the lower the propulsion that the wheel can transmit. We can also see that the wheels tend to spin faster and more easily for a flat curve profile. Approximately the following formula describes the behavior up to slip of 30%

$$\kappa \approx \sqrt{\sigma} . \qquad (14.3)$$

The curves shown in Fig. 14.8 are influenced by three features:

- Tyres and stud deformations
- Soil deformations
- Gliding on the contact surface

14.2.2 Tyres

Tyre Structure

Figure 14.9 shows the general structure of a tyre. The tyre consists of a casing and a contact surface. The casing is also laterally covered by rubber material. The wire-wound core keeps the tyre inside the rim's bead. Nowadays, drive wheel tyres are equipped with a hose in most cases, as the driving power transmission would otherwise be restricted.

Consisting of several tissue layers, the casing provides the tyre with stability. According to the casing's construction, different tyre construction varieties can be divided into the following groups:

- Biased ply construction (Fig. 14.9b): The tissue layers run from one bead to the other with 45° staggering. This construction variety is relatively easy, making the tyres cheaper.
- Radial ply construction (Fig. 14.9c): In this tyre type, also called belted tyres, the tissue layers run radially from bead to bead. Around this inner layer, on the contact surface, there is an additional contact surface consisting of several tissue layers (the belt). Thus, the contact surface gains greater stability. Due to the soft casing, the tyre exhibits much greater spring deflection and thus a larger footprint than that produced by a diagonal tyre. Additionally, there is:
  - Better force transmission
  - Reduced soil pressure
  - Reduced roll resistance
  - Softer spring deflection of the tractor
  - Extended life of the contact surface

Force Transmission Between Tyre and Soil

Various forces are transmitted from the tyre to the soil. As already mentioned above, the material properties of the soil considerably affect the driving mechanics.

The highest pressure transmitted by the tyre to the soil, and thus the greatest degree of compaction, is always found at the center of the contact surface. However, the dimensions of this contact surface depends on the deformability of the ground. This means that the footprint increases with a softer soil, i. e., the contact surface pressure decreases. In any case, the integral of the pressure stays the same (Fig. 14.10).

The soil pressure arising from the machine's weight considerably influences the soil fertility, thus being of special importance for agricultural machines. It

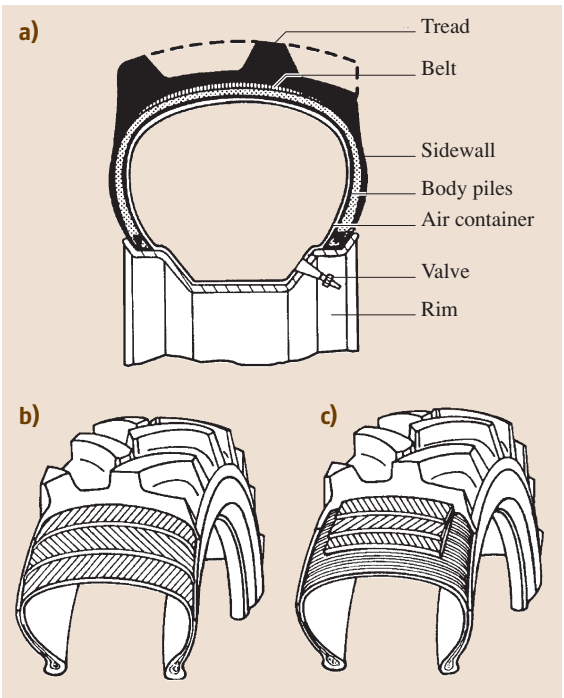


Fig. 14.9a–c Cross section through a tractor tyre with rim (after [14.10], description in text)

is important to minimize the negative impact of increased soil pressure by choosing an appropriate set of tyres. Due to internal friction in the soil, the pressure continuously decreases with distance from the footprint. However, the isobars, i. e., the lines of equal pressure, reach deeper into the soil with increasing load (Fig. 14.11). When the subsoiler does not work efficiently enough, detrimental soil compaction may occur.

The contact surface pressure is a decisive feature regarding the growth of new plants and soil ventilation. For solid ground, the following empirical rule apples: due to the casing's rigidity, the contact surface pressure on the footprint is about 0.03–0.04 MPa higher than the inner tyre pressure with radial tyres, and 0.05 MPa higher with diagonal tyres. For solid ground, this pressure is almost the same across the footprint because the deformation of the tyre is much greater than that of the soil [14.9].

In the case of earthmoving and road construction machinery, preserving soil fertility is less important. It is much more important to concentrate on features such as load-bearing capacity, insensitivity to stones, and good traction capacity. In the case of compaction rollers, it is precisely compaction which is important. In any case, it is necessary to consider these effects and their relations with the corresponding application.

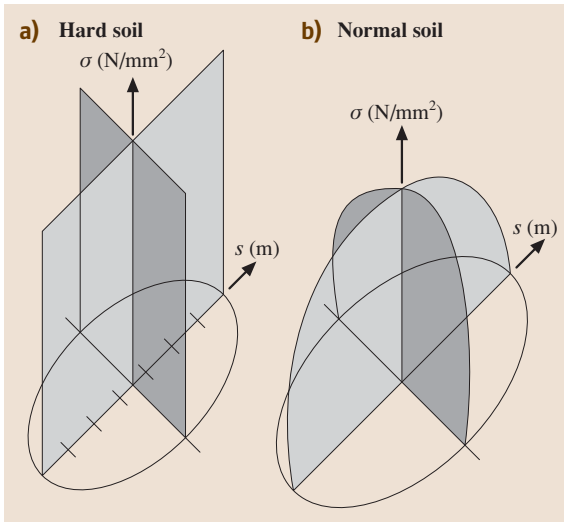


Fig. 14.10a,b Qualitative description of the pressure tension fields in the soil below the tyre (after [14.11])

Chain Chassis

Apart from tyre chassis, earthmoving and road construction machines mainly use a chain or caterpillar chassis. A caterpillar track consists of individual links arranged in front of the machine following the driving direction, thus serving as a lane for the machine. The chain has to simultaneously take up all forces resulting from traction and the load of the machine itself as well as possible lateral forces caused by steering movements. Similar to the tyre–soil system, the dynamic traction ratio between the chain and the soil depends on a number of parameters. Usually, the chain–soil system permits the transmission of large traction forces. Due to its footprint, which is larger than that of a wheel chassis,

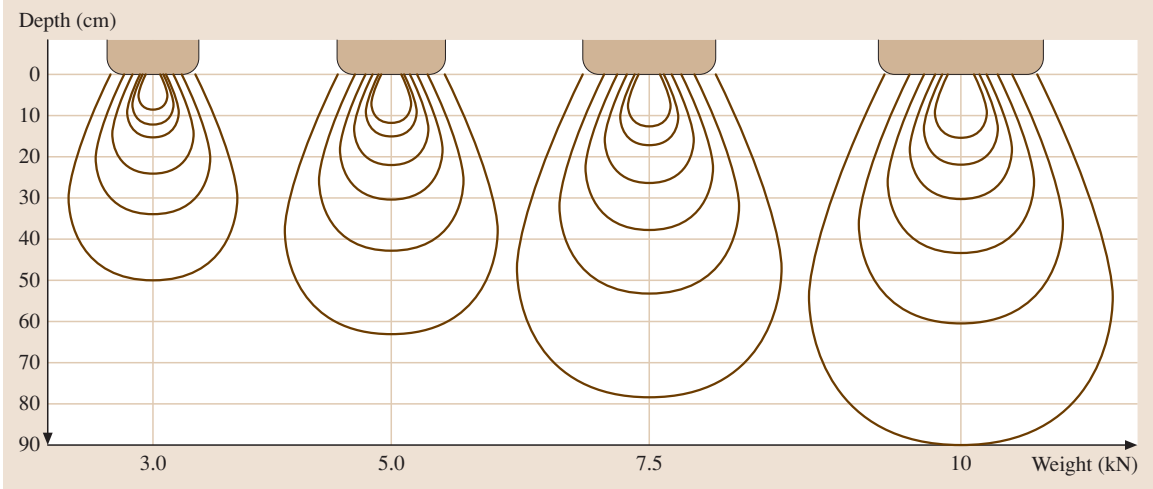


Fig. 14.11 Influence of the tyre load on the soil pressure (air pressure 0.082 MPa in all cases) [14.9]

the soil pressure is lower. Another feature of the chain chassis is its good stability, which is of considerable importance for hydraulic excavators.

Rubber belt tracks combine the advantages of both systems. They are applied in mini-excavators and road pavers. In the area of agriculture, rubber belt tracks are used in tractors equipped with high engine powers designed to work on the soil. Rubber belt tracks suffer from less vibrations compared with steel chains, while almost the same driving speeds are possible as with standard tractors. Moreover, they cause less damage to the road surface [14.12].

Figure 14.12 shows the general construction of a chain track. It consists of the chain, the base plate, the track rollers, the support rolls, the driving wheel, the guide wheel, the tensioning device, and the track frame. A chain chassis consists of at least two chain tracks. The guide wheel usually has only one guide profile and, using the chain tensioning device, can be shifted in order to tighten the chain. The smaller track rollers on the bottom serve to support the machine against the chain. The track rollers are designed to transmit even lateral forces emerging from steering movements or slope forces. There are various arrangements for the tractive turas. Whereas the conventional variety is preferred for hydraulic excavators and cold milling machines, large dozers are mainly equipped with the delta arrangement. Advantages of the latter include that the drive is protected against pollution and that driving performance is better.

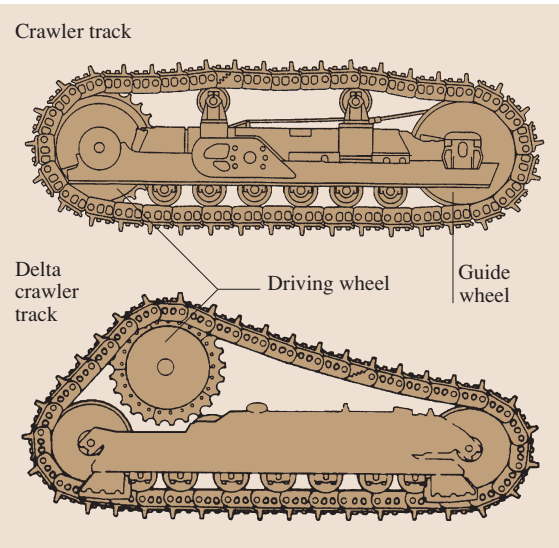


Fig. 14.12 General construction of a chain track

Steerings

The vehicle's chassis and the planned application are decisive for the choice of the steering system to be utilized. The steering considerably influences a machine's steering stability, driving security, stability, and manoeuvrability. In the case of tracklaying vehicles, the only useful steering method is skid steering. With a wheel chassis, vehicle steering is achieved by applying different driving speeds to the left and right chain tracks. A wheel chassis uses axle pivot steering of one or both axles, articulated steering or a combination of articulated and axle pivot steering. Whereas tractors mainly use axle pivot steering, earthmoving machines are equipped with the steering varieties shown in Fig. 14.13. The articulated steering variety is applied mainly in wheel loaders and is characterized by a two-part machine frame. An articulated chassis connects both frame elements, facilitating hydraulic displacement of the elements around the vertical axle. The steering angles are 40–50° to both sides. Given an additional movement around the vehicle's lateral axle, this articulation results in a center pendulum pivot steering system. Such an articulation replaces a pendulum axle. The characteristic features of the articulated steering are:

- Very good manoeuvrability
- The rear wheels are running in the same track as the front wheels
- Slim machine construction is possible
- When turning the wheels, the center of gravity is displaced, reducing the maximum load capacity

Axle pivot steering is an alternative to articulated steering. In most cases, tractors are equipped with axle pivot steering, on one or both axles. Various types are used in the case of all-wheel drive:

- Driving curves: Here, the wheels are turned so as to move the machine around the intersection point of the verticals or the prolonged axle. To achieve this, it is only necessary to steer one axle; in most cases, the front axle is used for steering.
- Four-wheel crab steering: The wheels are turned so as to move the vehicle laterally in a line parallel to its longitudinal axis.
- Circular driving: The wheels are turned so as to turn the machine on the spot.
- Diagonal driving: All four wheels are turned by 90° so as to move the vehicle diagonally to the vehicle's longitudinal axis.

The characteristic features of a vehicle equipped with axle pivot steering are:

- With axle pivot steering of the front axle, there are four tracks. which causes the driving resistance to increase on yielding soil.
- It is not possible to *free* the vehicle with only. one turn of the wheels as is possible with articulated steering.
- Good manoeuvrability is possible for all-wheel-driven vehicles.
- Good stability.
- Steering and turning the wheels requires space.
- Four-wheel crabbing facilitates lateral displacement of the machine.

Apart from these general types, there are also combinations of articulated steering and axle pivot steering.

14.2.3 Earthmoving Machinery

Earthmoving designates all modifications of the Earth's crust regarding position, form, and density. Among others, this might include:

- Excavate soil
- Digging ditches and soil excavations
- Raw material extraction
- Soil transport
- Material installation, e.g., for road and embankment construction
- Material compaction

Along with structural and civil engineering, earthmoving is another branch of construction, being applied in different cases.

Wheel Loaders

Wheel loaders or shovel dozers are extremely mobile machines with universal application possibilities. Their main task is to remove soil, as well as to load, transport, and locate material [14.13]. The working level is the same as the driving level, or even slightly below in some cases. Wheel loaders can be classified into:

- Front-end loaders
- Swing loaders
- Overhead loaders
- Telescopic loaders
- Skid steer loaders

The chassis and the steering variety used (articulated or axle pivot steering) are other criteria that can be used to distinguish between types of wheel loaders.

The most important variety of the wheel loaders listed above is the front-end loader. Here, the working device is mounted on the front part of the machine frame. It consists of a multiple bar linkage powered by hydraulic cylinders. This hoisting gear is usually equipped with a loader shovel. When loading soil, the shovel is lifted to the surface level. Keeping this position, the whole machine has to drive into the material. Thus, apart from the driving resistance, the machine also has to compensate the shovel's penetration into the material and filling resistance. When the shovel is filled, the hydraulic cylinders are placed into the transport position. Wheel loaders are mainly driven by diesel engines. Generally, a wheel loader has to be equipped with drives for the working devices, the steering units, and the chassis. Depending on the machine's power class, various drive varieties are used:

- *Wheel loaders of maximum* 59 kW. These wheel loaders mainly use hydrostatic drives. The efficiency factor, which is worse than that of mechanical drives, is compensated by lower costs and broader application. In general, the resulting maximum speed is 25 km/h and, in most cases, steering is by axle pivot.

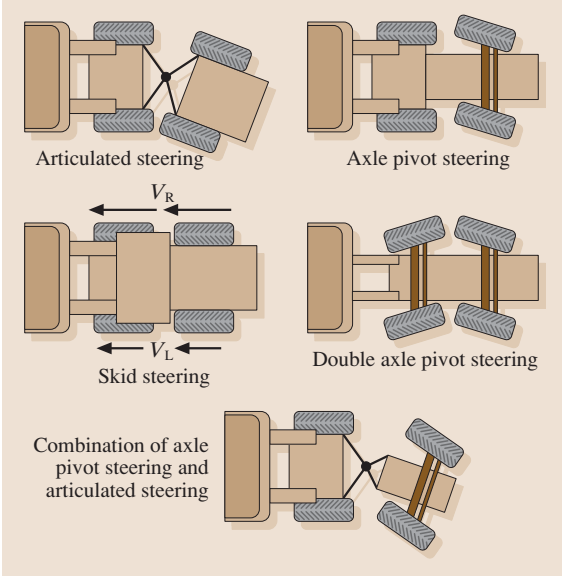


Fig. 14.13 Steering types

- *Wheel loaders of max.* 110 kW. The drive of these wheel loaders is realized by a power shift transmission and a torque converter. Some machines also use hydrostatic drives. Either articulated or axle pivot steering is applied. Maximum driving speed is about 40 km/h.
- *Wheel loaders of more than* 110 kW. Large wheel loaders are always equipped with power shift transmission, a torque converter, and articulated steering. Their maximum driving speed is 35–45 km/h. These machines often have a general operating license for general traffic.

Apart from driving power, there must be sufficient power for the working devices. Usually, the lifting force and the lifting speed of the working devices are adjusted, aiming to minimize the time period necessary for one loading process. Additionally, the lifting force is restricted to ensure stability of the machine. The working hydraulics usually consists of an open system with pilot control, which is used to supply the drives for steering and loading movements with an extra pump. Figure 14.14 shows an example of a hydrostatic drive of a wheel loader's working hydraulics.

The shovel's lifting and breaking off forces are mainly determined by the hydraulic cylinders and the kinematics applied. Figure 14.15 shows the kinematics used today. Modern machines are very often equipped with Z-kinematics due to the huge breaking off force effected by the good transmission of the shovel-tip cylinder mechanism. During the lifting process, Z-kinematics moves the shovel forwards and backwards while lowering the shovel. Parallel carrying is worse than with P-kinematics. The torque parallel (TP)-kinematic represents a compromise between both types of kinematics.

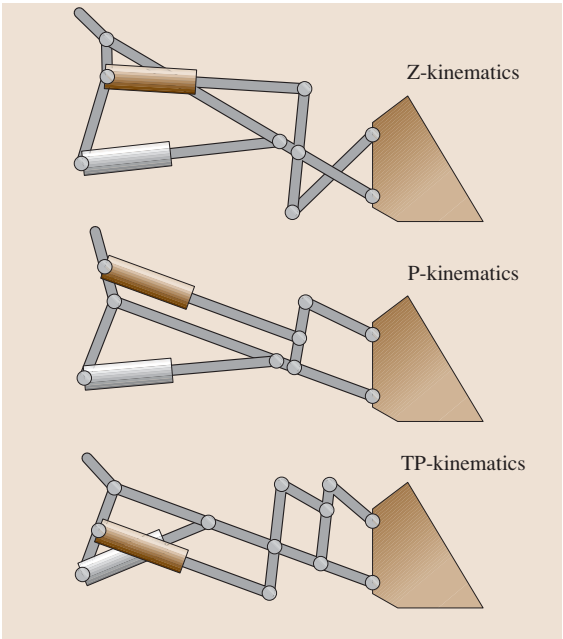


Fig. 14.15 Kinematics of a wheel loaders' working devices

Regarding the driver's working place, wheel loaders are similar to tractors. As a rule, these machines are equipped with a driver cabin today, not only for safety

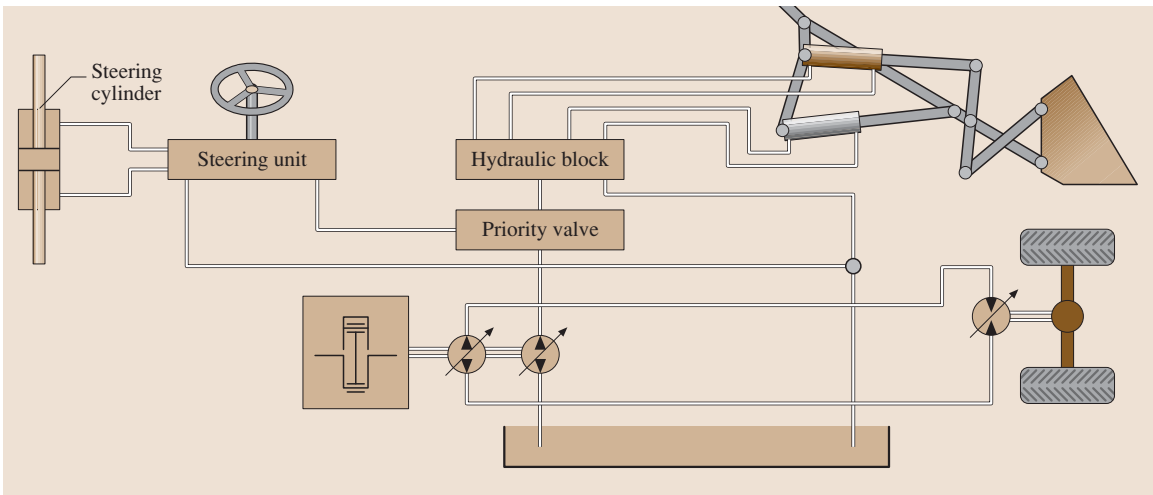


Fig. 14.14 Simplified hydraulic plan of a wheel loader

but also for driver comfort. Usually, the driver spends several hours operating the machine without a break. Therefore it is necessary to minimize stress and strain on the driver. The requirements for the driver's working place:

- Good view over the working device and the load during the whole working cycle
- Complete visibility without disturbances
- Huge door opening angle
- Safe and easy access to machine
- Air-conditioned cabin
- Adjustable driver's seat
- Ergonomic arrangement of the instruments
- Good vibration damping and reduction

The control system is also very similar to that of tractors. A controller area network (CAN) bus system is used for communication between the engine, operating device, and other machine components [14.14] (Fig. 14.16). The electronic components must be adjusted to rough machine usage, i. e., they must be well protected against vibrations, water, and dust. Nowadays, there are even microcontrollers that facilitate automated movement processes as well as supervising individual system states in order to prevent accidents.

Excavators

Excavators are loaders equipped with one or more buckets. A machine equipped with one bucket can only work discontinuously, i. e., carrying out intermittent process steps. In contrast to these *one-bucket excavators*, machines equipped with several buckets can work continuously.

Below we give a more detailed description of the so-called *one-bucket excavator* or cable dredger. This is an extremely large group, extending from special-purpose machines to universally applicable devices, so-called universal excavators.

The application area of the *one-bucket excavator* (or simply *excavator*) is extraction and handling of any kinds of goods. For this purpose, they use excavator spoons as well as digging and loading devices. Due to the advantages of hydraulic drive systems, excavators using this system are more important than those using a wire drive. The latter, however, are applied in excavators that are used to dig boreholes, which requires considerable working depth.

Today, the hydraulic excavator is the most important machine of this type, as hydrostatic power transmission and conversion are very advantageous in the case of excavators. Hydraulic excavators can be divided into single-purpose machines equipped with only one working device, and universal excavators, which can be equipped with various types of working devices [14.15]. There is a wide variety of such machines. Just regarding their own weight, they range from 0.5 t up to 800 t in the case of hydraulic excavators used in earthwork and open-cast mining. The smaller machines of 0.56 t are often called mini-excavators.

Another parameter used to classify these machines is chassis type:

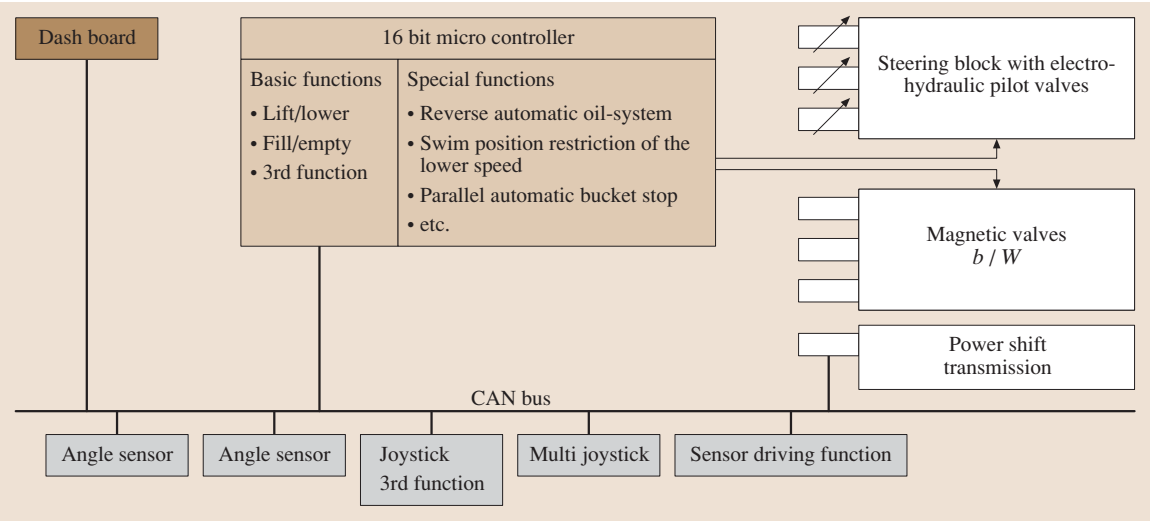


Fig. 14.16 Control system of a wheel loader (after [14.14])

- Tracklaying excavators, with crawlers or rubber-belted tracks
- Mobile excavators, which are equipped with a wheel chassis

Tracklaying excavators are designed for low driving speeds of up to 6 km/h. Normally, they just move on construction sites and do not participate in normal traffic. Mobile excavators or wheel excavators, however, do take part in the normal traffic. As is the case in agriculture, driving speeds are increasing considerably, with some machines traveling with a maximum speed of 50 km/h.

Figure 14.17 shows the basic construction of a hydraulic excavator. It is equipped with a revolving superstructure hinged to the chassis, which may be of chain or wheel type. The revolving superstructure carries the drive, including the oil and fuel containers as well as the cooling system, the filter, the control valve, the actuators, the driver's cabin, and the counterweight. The arrangement of these components is determined by the demand for balanced mass distribution, visibility, and ergonomics.

Another important component is the rotary transmission leadthrough, i. e., the connection between the revolving superstructure and the undercarriage. The hydraulic power needed for the drive is transmitted to the undercarriage by this rotary transmission leadthrough, which is a rotary pipe connection used to transport the fluid power from the rotating revolving superstructure to the drives installed in the undercarriage.

In the case of excavators, the hydraulic system is one of the most important components as all of the drives on the excavator are run hydrostatically. Using a pump power divider, a diesel engine distributes the power to several hydraulic cycles. There are always several power consumers to be supplied, which is a characteristic of excavator application. As a consequence, it is necessary to use control systems facilitating sensitive power-consumption adjustment and good energetic conditions as well.

Nowadays, an excavator's diesel engine is usually run at a rated speed, i. e., there is a constant initial speed of the hydraulic pumps. In some excavators the engine runs at a set of constant initial speeds. Variable-displacement pumps are used to adjust the flow rate depending on individual requirements. To achieve this, various systems can be applied, based on the load-sensing principle.

Hydraulic excavators increasingly make use of electrohydraulic systems. Microcontrollers also enable new possibilities for steering hydraulic devices and controlling and automating working cycles; for example, this kind of steering facilitates teach-in steering of individual motion sequences and automation of individual working cycles. Furthermore, there are even new possibilities regarding excavator management which do not involve running the diesel engine at a constant speed. In these systems, the diesel engine is optimally adjusted depending on individual requirements, which helps to reduce exhaust gas emissions and fuel consumption [14.17, 18].

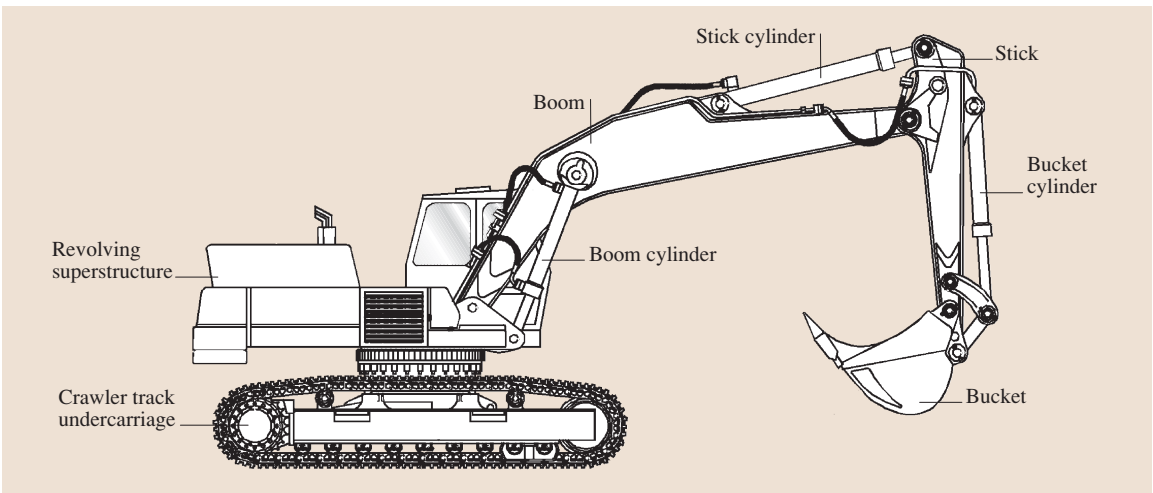


Fig. 14.17 Basic construction of a hydraulic excavator (after [14.16])

Graders and Scrapers

Graders are machines applied mainly in road construction (Fig. 14.18). The working device used is a blade designed for various applications. It is flexibly mounted between the front and the rear axle. In this position, the blade exhibits less vertical movement when crossing obstacles. In most cases, the chassis consists of pneumatic tyres. Apart from biaxial vehicles, triaxial ones also exist, with two axles combined into a tandem axle. In this case the single axle is used for steering, but machines with four-wheel steering or articulated steering have also been designed.

Graders are driven by a diesel engine with a capacity of 30–210 kW. Usually, the engine drives the tandem axle. Driving speed is 2–40 km/h, and can be accurately adjusted to the requirements of the individual task by using a driving gearbox. This is done through a manual transmission with up to 12 forward and 4 reverse gears. A grader's working weight is 5–30 t, and the wheelbase is 4.5 and 7 m. There are, however, devices that are much larger, weighing more than 90 t and with an engine power of more than 500 kW. For the vehicle's drive, various mechanical transmission types can be used, some of which have hydrodynamic converters or hydrostatic drives. Due to their advantages regarding various arrangement possibilities and considerable converting range, hydrostatic transmission has become increasing important in these machines. A decisive factor for the vehicle's drive is the nominal thrust force, which determines the machine's power capacity. The nominal thrust force depends on the engine power and the machine weight. A 2×2×2 grader delivers a ratio of engine power to machine weight of about 9–10 kW/t. A 1×2×3 grader is expected to deliver a ratio of about 7–8 kW/t [14.19].

Figure 14.19 shows a grader's blade positions. The specialized kinematics facilitate a large variety of blade positions, making the grader a universally applicable machine.

Depending on the chassis type, there are two types of scrapers: the wheel scraper and the tracklaying scraper with crawler, which is mainly utilized for difficult soil conditions. The excavating bucket consists of several parts. It has movable front and back sides in order to pour the material out of the container in a particular direction. Depending on the material and the machine, the cutting depth is 5–200 mm. The container's volume may vary from 1 up to 40 m^3, and the engine power may be as high as 700 kW. The driving speed necessary to transport the material can be 10–50 km/h. Conceptually, the front wheels are most strongly affected by the load, so they are driven. However, four-wheel drives are also in use [14.19].

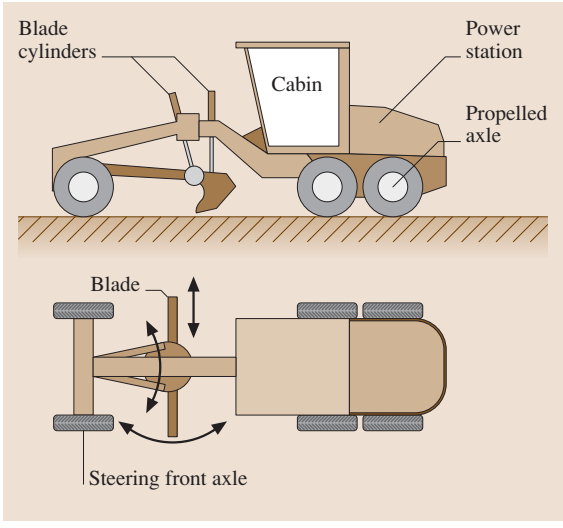


Fig. 14.18 Grader

14.2.4 Road Construction Machinery

Compaction Machinery

Compaction aims to increase the storage density and reduce the pore volume of material. Compaction machines serve to crush cavities and store soil grains at the highest possible density. The methods used depend particularly on the soil type:

- Grit and sand are compacted by vibration or pushing. Material of the same grain size is usefully mixed with other material in order to obtain a sieve classification. Pores between the coarse grains

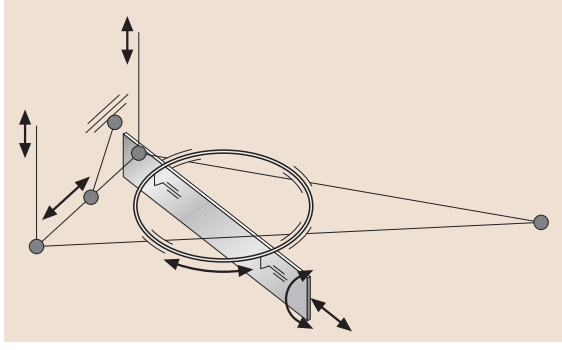


Fig. 14.19 Blade kinematics of a grader

should be filled with finer grains. The devices used are: vibration plates, vibration rollers, and blasting.
- Gravel and detrital rocks are compacted by pounding or pushing, smashing bulky chunks. The devices used are impact stampers, explosions stampers.
- Silt and clay are compacted by rolling or dispersing. There has to be sufficient water content in order to reduce the soil friction. However, the water content should not be lower than the pore volume to be achieved by compaction, otherwise water saturation will occur and the soil will adopt a plastic consistency. The devices in use are rammer butt rollers.

The compaction process consists of several individual processes:

- Ordering or distribution
- Destruction
- Pore water transport
- Displacement

Two main machinery groups can be distinguished:

- Statically working machines
- Dynamically working machines

It is also possible to classify them into rollers and vibratory plates.

Statically working rollers are among the oldest compaction devices, and today they exist in a huge variety. Depending on the number of roll bodies, they are classified into one-, two-, and trial-axle devices, or depending on their arrangement, into tandem or tricycle rollers. The roll body itself may be constructed as a flat roller, equipped with so-called sheep-feet roller, as a grid roller, or having rubber wheels.

Today, earthmoving and road construction make use of dynamically working rollers due to their better compaction performance. Usually, the rollers are equipped with vibratory stimulation; in some machines, there oscillation stimulation is also provided (Fig. 14.20). The unbalanced arrangement of masses and the rotation facilitate the introduction of vertical and rotating swinging into the drum or roller. The vertical swinging corresponds to vibration, and the rotating swinging provokes oscillation. By combining both movements it is possible to work on the material to be compacted according to its compaction potential [14.20]. Special measuring and steering systems enable optimum compaction. In these systems, there is a special focus on measuring the compaction. Modern systems are based on the principle of acceleration sensors that continuously measure the increasing rebound acceleration, comparing it with the result of previous measurements. In the case of the detection of increased compaction, the sensor produces a signal to carry out another drive over. For future documentation, the measurement data are additionally stored in relation as a function of area or position [14.21, 22].

With the vibration stimulation set to make the roller jump, the layer to be compacted will be more or less destroyed as it becomes uneven [14.21, 22].

The forerunner of the stamper was the iron hand stamper, which was used for paving work. In order to increase its stamping effect, engineers have developed heavier devices that are lifted either by excavators (the freefall stamper) or by blasting. The explosion stamper is used for compaction in narrow excavations or shafts. By igniting a fuel mixture, a rammer is thrown up (up to 46 cm), and then falls to the ground under the effect of gravity. This process provokes compaction of the fuel and re-ignition. Compaction of the soil is then effected by the explosion pressure and the rammer's rebound over a large number of strokes. A small stamper weighing up to 100 kg delivers compaction depths of up to 50 cm; heavier devices (500–1200 kg)

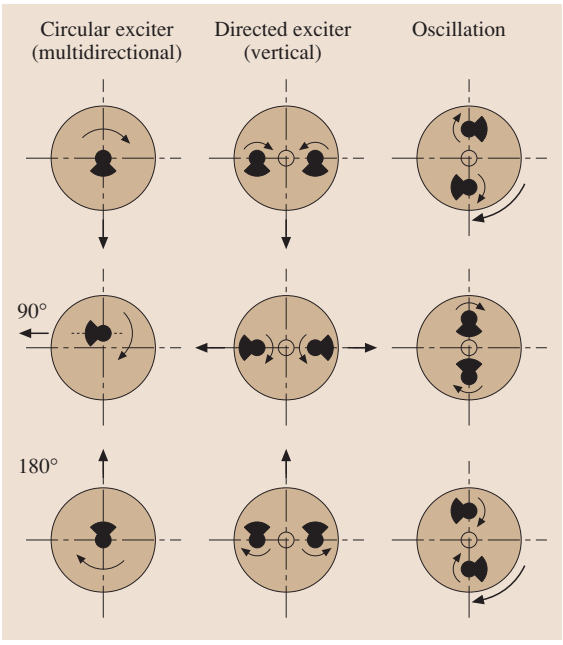


Fig. 14.20 The principle of vibration and oscillation stimulation (after [14.20])

achieve depths of 40–90 cm. Depending on the device used, the number of strokes is 60–80 strokes per minute. Vibratory stampers are used in narrow locations just like the explosion stamper. However, with a weight of 60–96 kg, they are lighter and only jump by 3–8 cm. Their multiple spring system is driven by combustion or electric motors with an average capacity of 2–3 kW. The material is compacted by the huge number of strokes (400–1000 strokes per minute) which cause a riddling effect. The vibratory plates are driven by combustion engines or electric motors. The vibrations are caused by an unbalanced mass vibration generator of various types of construction (directed and undirected vibration). Smaller plates have to be pulled to and fro, whereas larger devices are automotive, being steered by a shaft. It is possible to control their work direction, speed, and vibration.

Machines Used for New Road Construction

A variety of machines are used for road construction, and they shall be described in detail below. Figure 14.21 provides a rough classification of road construction machinery related to the individual construction processes.

Road pavers and slip-form pavers are used to build new road layers. Whereas road pavers are used to install special types of concrete, slip-form pavers are exclusively utilized to lay concrete. When laying asphalt, it is necessary to carry out a final treatment with rollers. When installing concrete, this roller work is not necessary. The slip-form paver compacts the material directly inside the machine, using special elements. In order to achieve the desired surface structure of the road, special-purpose machines are often used to work on the material when the slip-form paver has passed over it.

Road Pavers. The actual laying device is a screed. The road paver's screed is swimming on the material, and thus is also called a *swimming screed*. The buoyancy conditions depend on the weight, the screed's angle, the forward speed, and the material's viscosity.

Fig. 14.21 Process chain of road new construction

By means of tension bars, the screed is connected to the vehicle. The connecting spot is situated approximately in the center of the chassis to ensure that the vehicle's pitch does not affect the screed. If the layer thickness is to be changed, one also has to change the height of the linking point of the tension bars. Due to the swimming principle used in these machines, slow material alteration is created, leading to a long-wavelength height variation which is good for driving comfort. Nowadays, steering is exclusively done by electronic leveling systems, which facilitate separate steering of the left and right linkage point of the tension bar.

The screed itself is equipped with additional compaction devices. Using a tamper, the material is stuffed in front of the screed and compacted. For additional density increase, vibration elements or pressure bars used. In order to produce good installation of the hot material, it is necessary to heat the screed's compaction devices that are in contact with the material. This is done by electric systems powered by a generator that is run by a diesel engine. Furthermore gas and liquid heating systems are used for heating the screed using burners. The working width of road pavers is 1–16 m, with deposition layers of up to 0.35 m being possible. Given appropriate material logistics, large road pavers can install more than 1000 t of material per hour [14.21] (Fig. 14.22).

The paver tractor consists of a frame to which the chassis is mounted. Furthermore, it contains the driving unit, consisting of a diesel engine, the gearbox, the cooler, and the hydraulic system. The driver stand is situated in the upper part of the vehicle, giving the driver a good view of the material container and the screed. The chassis is either a wheel or a caterpillar chassis. The caterpillar chassis is composed of two hydrostatically driven chain tracks which are electronically controlled. Apart from steel chains, rubber bands are also often used. The wheel chassis are often triaxial with a large rear-wheel drive and small wheels mounted under the material container. Usually, two axles are driven, but four-wheel drives are also used. The advantage of a wheel chassis is that driving speed of up to 20 km/h are possible, whereas a crawler chassis only permits driving speeds of up to 6 km/h, and caterpillar chassis equipped with rubber belts enable maximum driving speeds of 16 km/h. However, the crawler chassis offers better traction and is less sen-

sitive to soil unevenness, facilitating more even road construction.

In order to meet the requirements of an even layer density and of very even roads, it is necessary to have a regular work process without interruptions. This is only possible with an automotive material reservoir to transport the material from the trucks to the paver in a regulated way [14.23].

In many cases, factors associated with the construction, e.g., too small construction sites, why the paver must work without such a feeder. For this reason, the traction drive as well as all drives for material transport have to be controllable. The command variable of the individual control cycles is the amount of material. In order to measure the amounts of material in the corresponding transport areas, either mechanical material sensors or ultrasound distance sensors are used, with the measured values being processed by digital steering systems.

Compacting the asphalt layer is necessary to ensure its stability. Immediate compaction is carried out by the paver's screed. The following rollers are then used to obtain the final density. The following compaction elements are in use:

- Tampers: The tamper is the first compaction device; it is a tamper bar driven by an excenter shaft. The throw is determined by the excenter radius and can be gradually adjusted. Depending on the material, the throw is about 2–7 mm. Another influencing factor is the drive speed, which ranges from 600 to 2400 rpm.
- Double tampers: Another possibility is to use a double tamper, which is driven by an excenter.
- Vibration stimulator: The majority of screeds make use of the vibration principle in order to compact the layers. Usually, a vibration stimulator is mounted on the blade of the screed, being driven mechanically with an adjustable frequency.
- Pressure bars: Pressure bars are similar to tampers. They are driven by pressure impulses which are transmitted into the layer by the bar, thus effecting compaction.

To meet the required geometrical dimension of the road to be constructed, use is made of leveling systems. Graders, dozers, slip-form pavers, and milling machines are also used, designed to fulfil the following tasks [14.21, 24]:

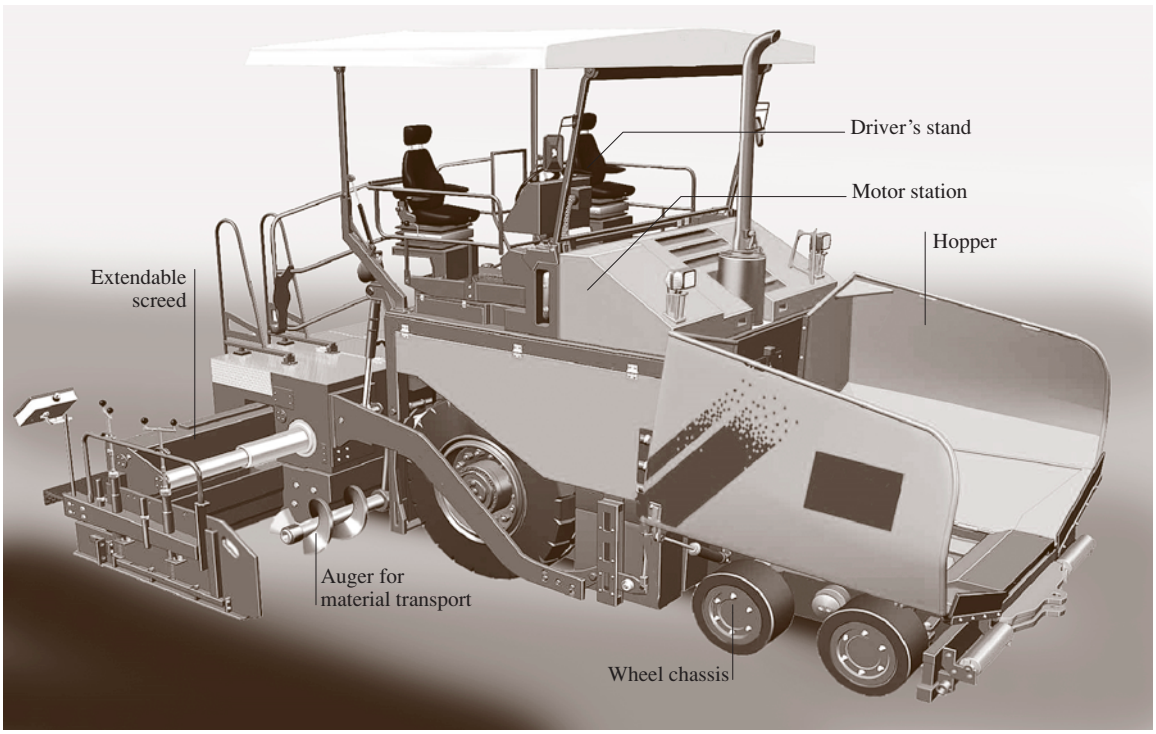


Fig. 14.22 Road pavers

- Position control of the working device
- Height and inclination control of the working device
- Position control of the working device

In the case of road pavers leveling is carried out by height adjustment of the linking points of the tension bar.

Slip-Form Pavers. For concrete road pavement, mainly slip-form pavers are in use today (Fig. 14.23). These machines are characterized by:

- Dragging along the *slipping form*
- Series arrangement of the working devices:
  - To compact the delivered concrete
  - To compact the concrete using vibrators
  - To pave
  - To fix anchors
  - To screed

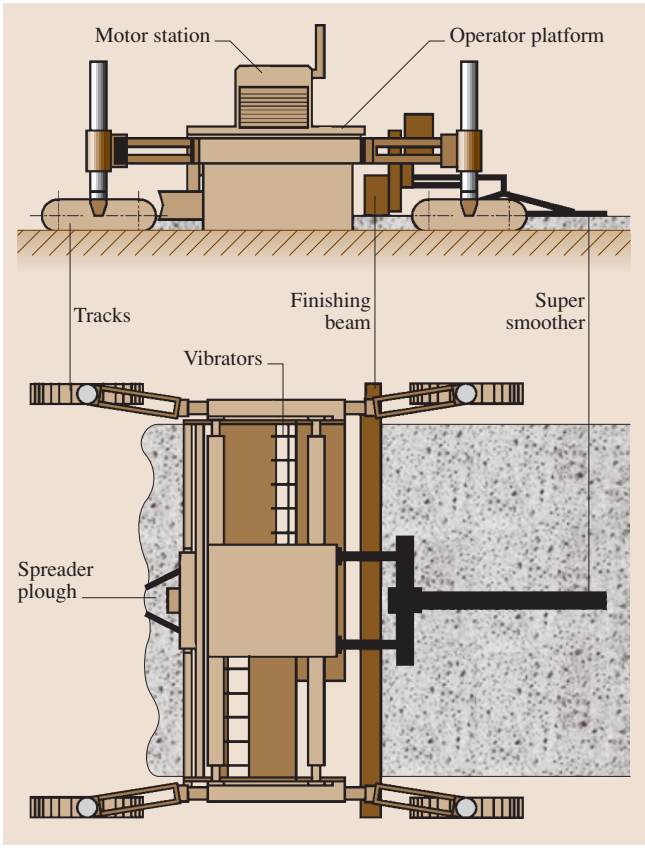


Fig. 14.23 Slip-form paver with variable working widths (after [14.25])

The working devices are connected to the tracklaying chassis in a frame. In contrast to road pavers, leveling in these devices is carried out by lifting and lowering the machine frame. To achieve this, the frame is connected to the tracklaying chassis by hydrostatically driven hydraulic cylinders. Similar to the road pavers, a diesel engine is used for the drive, providing the hydraulic and the electrical system with the energy needed. The installed engine power is about 79 kW in smaller machines, and in larger ones can be up to 300 kW. In smaller machines the chassis consists of three tracks, whereas in larger ones there are four chain tracks. The maximum working width is 16 m, and the layer thickness can be up to 0.45 m [14.21].

Machines Designed for Road Maintenance

Ageing of road materials as well as the weight permanently put on the coating require maintenance measures to be taken in order to ensure the road's function. Machine application depends on the particular road damage and the size of the coat to be worked on. There are five different paving processes [14.26]:

Reshape:	Reshaping of a road layer without new mixing material
Regrip:	Rebuilding of the road grip
Repave:	Rebuilding of a road layer with new mixing material
Remix:	Rebuilding of a road layer with milled and new material
Remix Plus:	Manipulating of the asphalt mixture in combination of rebuilding a road layer

The remix process usually is carried out in situ, i. e., the individual process parts of taking off, mixing, adding new material, laying, and compacting are done in only one stage. Repaving is usually an intermittent process, i. e., after taking off the material (e.g., by cold-milling machines) there is a delayed laying of the material with a paver followed by compaction by rollers.

The focus of the road maintenance process is taking off the damaged material. In the case of asphalt and partially for concrete (Fig. 14.24), this is done by cold milling machines. Here, a diesel engine drives the milling drum using a belt transmission. The milling drum is equipped with cylindric chisels, which serve as cutting instruments, situated in replaceable chisel holders. The roller's equipment as well as the size and the variety of chisels are determined by the application area. In order to avoid dust, the work field is sprayed with water. The removed material is transmit-

ted to a belt conveyor from where it is transported to a truck.

Small milling machines mainly use wheel chassis with three or four wheels. Their working widths are 0.25–1.2 m. Large milling machines, however, offer working widths of up to 2.2 m, with an engine power of about 600 kW. Depending on the machine sizes, milling machines are equipped with wheel or caterpillar tracks which are driven by hydro-motors (Fig. 14.25). The wheels as well as the chain tracks are mounted on lifting cylinders which are connected to the machine's chassis. The milling depth is adjusted by lifting or lowering the machine, controlled by an electronic leveling system.

A machine which is very similar to the milling machine is the recycler (Figs. 14.26, 14.27). It takes off the damaged material and recycles it. Very often various substances or binders such as foamed bitumen, cement or bitumen are added to the material which is to be prepared. Instead of a milling rotor as is used in cold-milling machines, a rotor equipped with cylindric chisels is designed for milling and mixing the milled material with the additional substances.

These vehicles use wheel chassis and tracklaying chassis. Their working widths are up to 3.05 m, working depth is up to 0.5 m, and the installed engine power reaches 500 kW.

The complex processes of in situ road maintenance (shown in Fig. 14.24) requires the machines shown in Fig. 14.27. This machine integrates the work functions of milling, breaking, taking off, mixing and adding new material, spreading, carrying them out in one stage, and precompacting. This helps to save time and reduce traffic obstructions. A machine such as this delivers working widths of 4.2 m.

14.2.5 Farming Equipment

Agriculture's main task is to produce sufficient food to nourish man and farm animals. Since the early Stone Age, when man began to settle and grow plants and animals, people have developed useful devices and

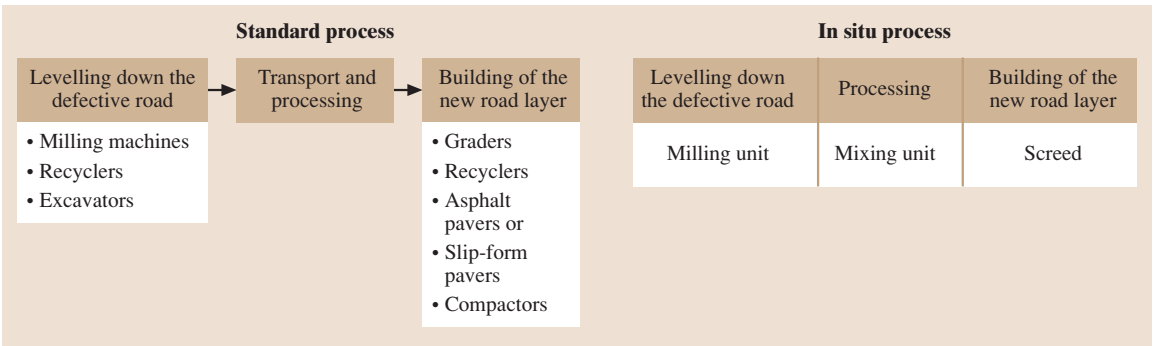


Fig. 14.24 Process chain of road maintenance

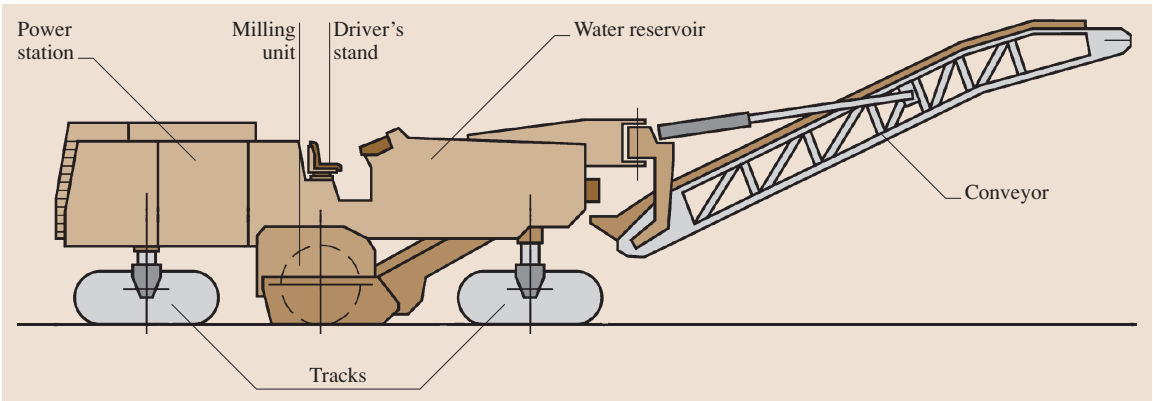


Fig. 14.25 Structure of a cold-milling machine

machines to do the necessary work. Mechanization, however, did not start until the 19th century, when steam engines, which could also drive larger devices such as the steam plough, emerged. Mechanization took a great leap forward in the 1950s when the agricultural sector lost increasing numbers of workers and new production processes were introduced.

The basic tasks of agricultural mechanization are:

- Improving productivity
- Increasing plant and animal yield
- Loss reduction
- Improving work productivity
- Reducing working hours
- Improving work process efficiency
- Improving workers' conditions

These aims lead to greater productivity and reduced resource consumption. Thus, agriculture has the same basic goals as industrial production, but of different goods. The decisive difference is the fact that agriculture is highly dependent on external conditions which cannot be influenced, such as weather and soil conditions. Due to the fact that agriculture is closely related to nature, it is necessary to consider ecological aspects as well.

Classification of Agricultural Machines

In general, agricultural machines are classified into machines used for livestock farming and those used for cultivation. Livestock farming deals with the production of animal products with the corresponding buildings, milking systems, etc. Machines from the mobile working sector are used for outdoor farming, as described in detail below. Due to the large variety of cultivated plants and considerable regional differences, numerous different agricultural production methods and machines have emerged. These machines may be classified as follows:

- *Tillage*: Tillage aims to change soil resources and conditions in order to provide plants with optimum growing conditions.
- *Sowing and planting*: These machines are designed to introduce the seed or the plants into the soil. Due

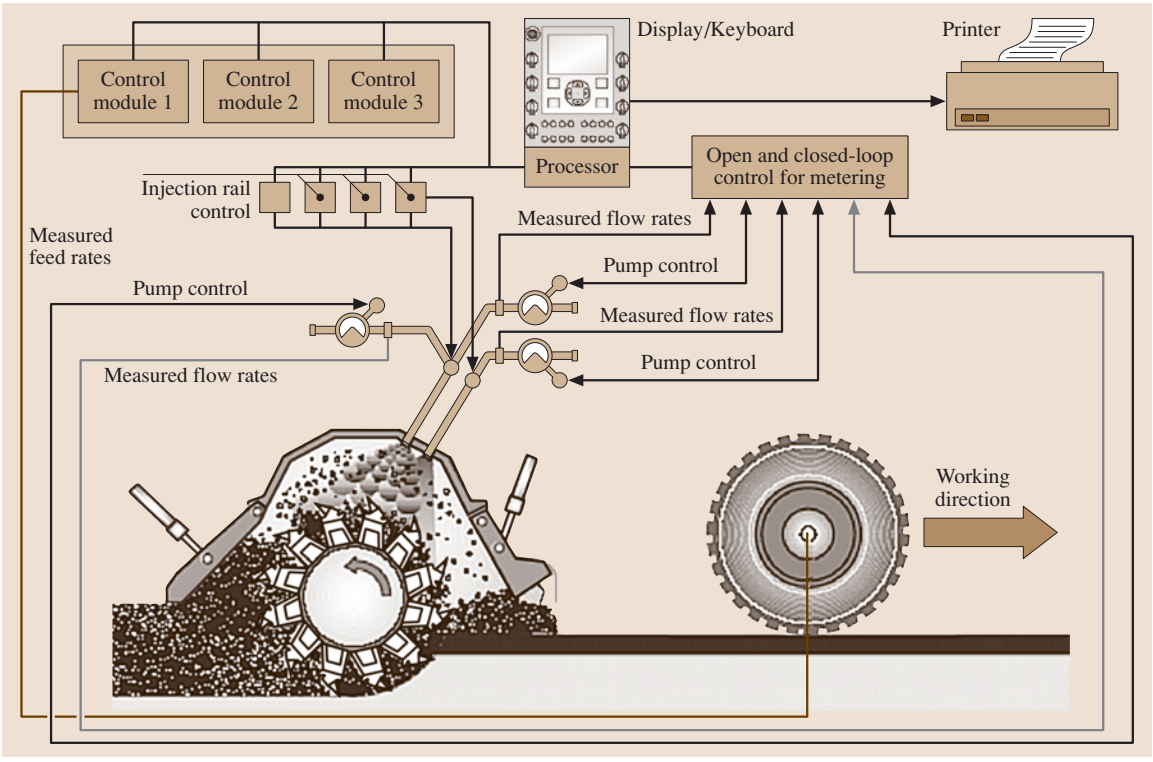


Fig. 14.26 Schematic diagram of the microprocessor control for simultaneous injection of foamed bitumen and water in a recycler (after [14.27])

to the large differences between plants, machine designs vary greatly:

- Drilling machines: Using dose-measuring devices and seed pipes, drilling machines are used to sow seeds. Usually seeds are sown in rows, with their distribution being as regular as possible and continuously adjustable.
- Single-seed drilling machine: This machine lays a single seed at a particular soil location. This is necessary, e.g., with corn and sugar beets, which need to be planted at a specified separation to ensure successful growth.
- Planting machine: These are special machines designed to lay tubers, e.g., potatoes, into the seed bed. They should fulfil the following requirements: regular, constant planting depth, exact tuber distance even with different tuber sizes and shapes, and regular row distance with even soil covering of the tubers.
- Transplanters: They are for plants, e.g., trees or certain sorts of vegetable.

- *Fertilizing*: Fertilizing machines are designed to distribute fertilizers regularly onto fields. There are two main groups of fertilizers:
  - Organic fertilizers (e.g., solid and fluid manure)
  - Mineral fertilizers
- *Plant protection*: The aim of plant protection is to protect crop plants against damage effected by weeds, fungi, and diseases. To achieve this, various methods are employed:
  - Physical-mechanical methods
  - Chemical methods
  - Biological methods

Mechanical plant protection removes weeds mechanically. Chemical and biological plant protection make use of plant-protecting agents. The machines used for this purpose are sprayers that mix fluid plant-protection agents and distribute them by using a hydraulic pump system equipped with spray valves. Modern sensor systems are able to identify damaged plants so that the agents can be distributed as required. Fruit culture and viniculture additionally make use of vaporizers, which – in contrast to sprayers – use air to support the drop transport.

- *Crop harvesting*: The majority of agricultural goods are crops such as grasses or cereals. Harvesting these culture plants means mowing their stems. The cutting method depends on several factors, the most important of which are:
  - Crop humidity
  - Crop geometry and stability
  - Harvest aim

  After the cutting process, the crop is further processed. In the case of grass, conditioners are used to destroy the external blade strata in order to cause the humidity to leave the plants faster, thus accelerating the drying process. Hay treatment machines are used to turn the grass so that it dries faster. They are additionally used to gather the crop into swaths so that a forage harvester or compactor can pick it up subsequently. It is the compactor's task to compact the crop, thus reducing its transport volume. Crop choppers are designed to chop the material so that it can be processed to become silage.
- *Grain harvesting*: Grain is one of the most important agricultural plant to be cultivated; it is

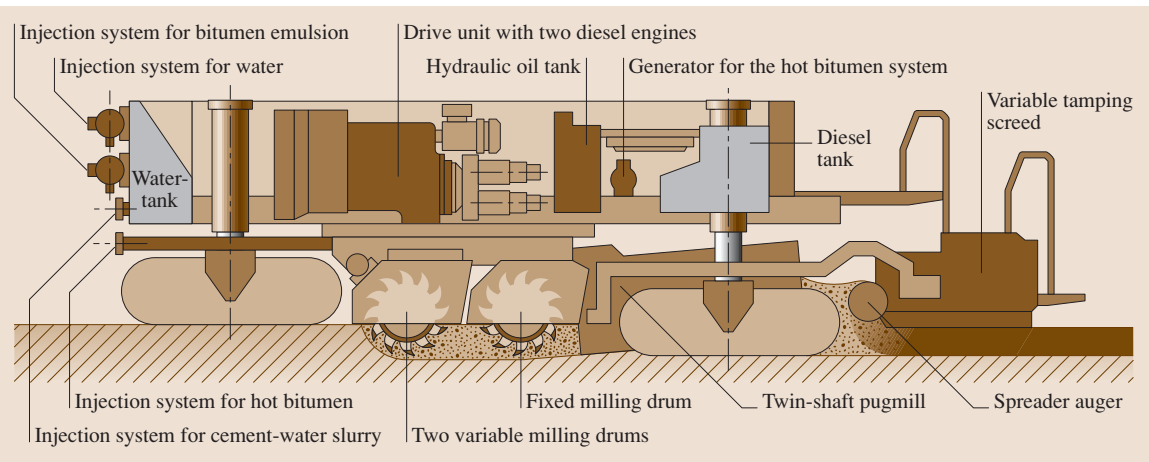


Fig. 14.27 Configuration of an in situ recycler (after [14.27])

harvested by combine harvesters. These machines are extremely complex, applying several mechanical process technology features:
- Cutting stems and ears
- Conveying stems and ears
- Threshing
- Separating grains from straw
- Cleaning
- Temporary crop storage
- Chopping straw

- *Root crop harvesting*: Potatoes and sugar beets are among the most important root crops. Harvesting them involves digging the crops out of the soil, cleaning, and then transporting them from the field. In terms of complexity, machines for this purpose can be compared to combine harvesters.
- *Engineering for intensive cropping*: The machines used for this task are special-purpose machines such as cotton harvesters or grapes harvesters.

Tractors

In the area of agriculture, tractors are of special importance as half of all agricultural machine investment is for tractors [14.9]. The tractor's tasks are [14.10]:

- Traction work on the field
- Transport work on the farm
- Driving mobile and stationary machines and devices
- Working with mounted devices

Figure 14.28 shows the most important tractor constructions. Most of the tractors used today are standard tractors. If one looks at the specified construction types, the universal tractor (standard tractor), equipped with a comfortable cabin and four-wheel drive, is of central importance. The huge range of different tractor concepts is reflected in the installed engine power capacity, which may vary from a few to several hundred kW.

Frames

The frame is the actual undercarriage of mobile working machines. In the case of tractors, the function of carrying is partially integrated in key components such as engine, gears, and axles so that there are different construction types:

- Block construction or frameless construction: This is the prevailing construction type used in standard tractors. Its characteristic feature is the link of axles, gears, and engine by means of flanges. The cases of these components are usually self-supporting cast constructions.
- Partial frame construction: This is a combination of block and frame construction, with various possible configurations. Here, the engine and the front axle are usually mounted on a steel frame. The gears and the rear axle, however, are still constructed as self-supporting components. In the case of the so-called *three-fourth frame*, only the rear axle is a self-supporting construction.

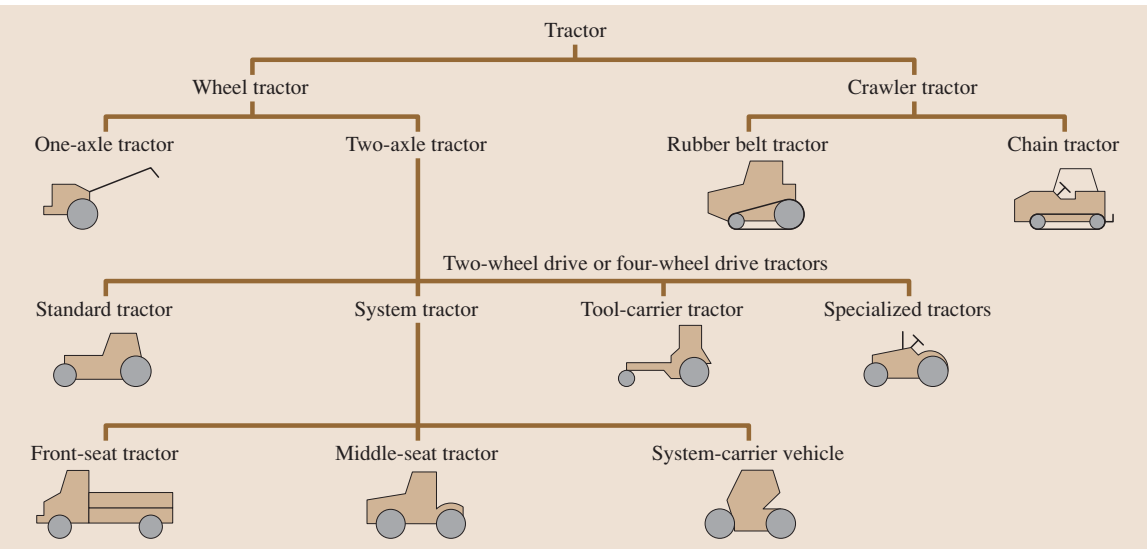


Fig. 14.28 Systematic classification of tractors (after [14.10, 28])

- Full-frame construction: Here the frame has to completely fulfil the task of the chassis. This approach is mostly used in machines with a small number of pieces, in system tractors, and in machines with a large range of engine and gear variants.

Chassis and Axles

In modern tractors, it is possible to drive all axles; pure rear-axle drives are the exception rather than the rule. The advantages of the four-wheel drive are:

- Larger traction force for field work (up to about 40% more for a traction weight increase of only 20%)
- Increased traction capacity, with equal slip and engine use
- Reduced energy consumption relative to tractive capacity
- Improved work with front devices, such as frontend mowers
- Improved driving safety
- Soil protection
- Smaller rear wheels
- Cost-saving front wheel brake

The disadvantages of the four-wheel drive are:

- Higher costs for the same engine capacity
- Greater losses, even with passive four-wheel drive
- Increased maintenance standards
- Larger turning circle

In the case of tractors, suspension and damping of the vehicle is done completely by the tyres. Due to the increased speed of standard tractors, up to 40–50 km/h, the suspension of the machine is of growing importance. As the rear axles are rigid, except for in a few exceptions, standard tractors are equipped with suspended front axles. Much use is now made of front-axle suspension with rigid axles, height control, and the possibility of locking the suspension for frontend loader work [14.29].

Power Transmissions

Tractors mainly use diesel engines due to their lower fuel consumption. This is achieved by direct diesel injection and the engine being charged by an exhaust gas turbocharger and air cooling. Over the last few years, there has been increased use of electronic engine controllers, facilitating directed control of the injected fuel and the injection time, which again helps to increase the engine capacity. Furthermore, it is necessary to combine modern injection systems and charging systems in order to keep in line with stricter exhaust gas regulations.

Tractor engine development is very closely related to that of other commercial vehicles today. The tractor's frameless construction impeded the use of standard engines whose case constructions could not take up the loads that would have been necessary for block construction. Meanwhile many tractors have a frame construction, easing this restriction.

Due to the complexity of the drive – the power take-off (p.t.o.) drive and four-wheel drive – a tractor's transmission makes production very complex. It is the gears' main task to adjust the tractor's driving speed to the individual working conditions. The tractor's rated power can only be transmitted with the rated torque. Therefore, the drive wheels' torque has to be adjusted in order to transmit the desired power capacity. Further tasks of the gears are changing the driving direction and driving auxiliary drives (p.t.o., four-wheel drive, frontend mower).

In standard tractors, the basic construction of the gears has been influenced by the frameless construction for a long time (Fig. 14.29). However, in recent

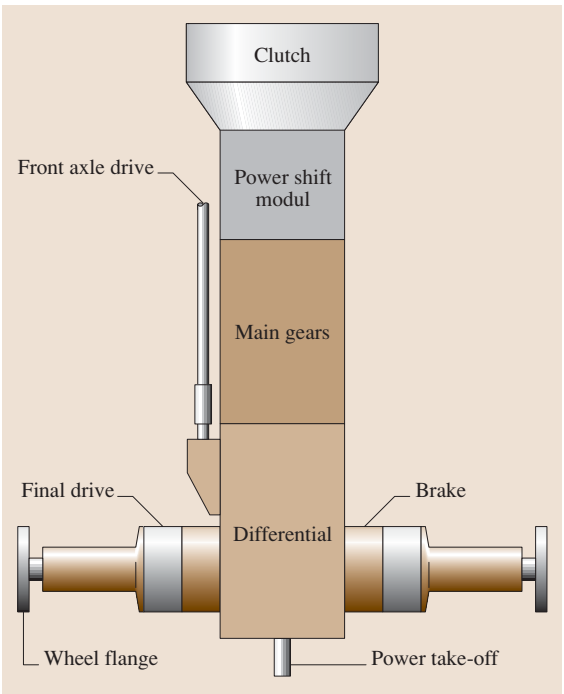


Fig. 14.29 Widespread arrangement of the most important functional groups in tractor gears of the main work pieces area (after [14.9])

years a tendency towards using frame constructions has emerged.

The engine transmits the power to the gears through the driving coupling and p.t.o. coupling. After fruitless attempts to establish direct conventional transmission in standard tractors until recent years, engineers mainly worked on power-shifted transmissions. Gear-shifted transmissions are often constructed according the so-called *group principle*. This variety of construction saves both gears and time. Due to the multiplication effect, a maximum of eight wheel pairs are needed; e.g., 16 forward steps with four basic gears and four groups. In practice, the groups are arranged in order to provide the driver with favorable work speeds (e.g., one field group "L" and one road group "H").

The latest result of tractor development is the continuously variable transmission. The advantages of this concept are good driver comfort, wide spread speeds, good efficiency, and flexible distribution of the load to the drive and p.t.o. In a continuously variable transmission, the power from the diesel engine is distributed by a planetary gear to a mechanical and a hydrostatic gear unit. Before leaving the gears, the changed power parts are added again. This results in good use of the mechanical drive's high efficiency as well as of the hydrostatic drive's shifted work process.

Figure 14.30 shows the basic construction of a tractor's continuously variable transmission using two gears. Due to the mechanical two gear transmission, which is applied only in the standard tractor, there are two driving speeds: 0–32 km/h and 0–50 km/h.

It is an important task of tractors to supply power to drive working devices. Apart from the p.t.o. delivering the mechanical power, tractors also supply hydraulic power. On the one hand, the latter is needed for driving working machines; on the other hand, hydraulics are also used to drive the hitch. Hydraulic systems for tractors have gained in importance. Today, there are two main tractor hydraulic systems to be distinguished in their basic construction:

- *Constant-flow system*. These simple hydraulic systems mainly use gear wheel pumps. The system's pressure depends on the required load, with the maximum pressure being determined by a pressure restriction valve. When a wheel gear pump is used, volume flow inside the system is only possible by varying the pump's rotation speed. The striking advantage of this system is its low cost. The disadvantage of these systems is that, if several devices are consuming power, they will influence one another. Furthermore, it is always the consumer with the smallest load that is supplied first. Another disadvantage is that the part of the volume flow that is not needed by a consumer has to be throttled away, which results in considerable losses. Modern constant-flow systems use multicycle systems running different consumers independently from one another by using separate pumps.
- *Load-sensing systems*. These are the modern hydraulic systems used in tractors, making use of axial piston displacement pumps. They allow continuous adjustment of the volume flow. A characteristic feature of a load-sensing system is that the volume flow is adjusted by a valve depending on a pressure drop. With the valve closed, only a small volume flow is provided to maintain a pressure of 2–3 MPa. When a valve is opened, the load pressure increases and is transmitted to the displacement pump's control by means of load pressure re-registration. The volume flow controller shifts the pump to a higher volume, thus keeping the pressure flow through the valve at a constant rate. The advantage of this system is that several power consumers can be supplied simultaneously with oil by one pump.

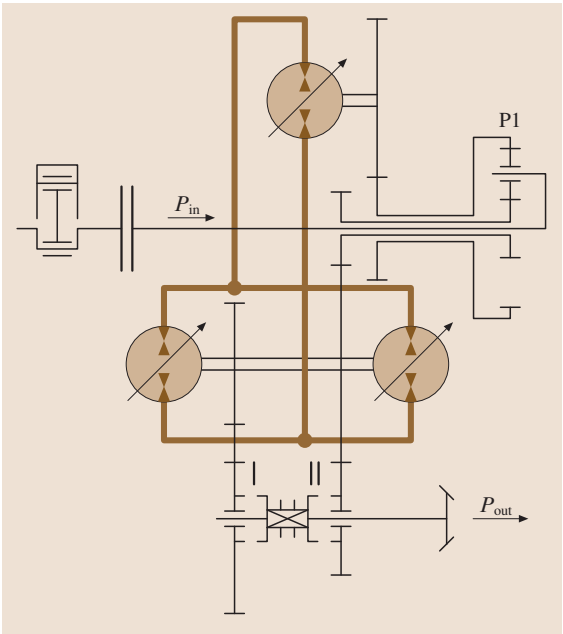


Fig. 14.30 Structure of the continuously variable transmission (after [14.30])

Operation Elements

It is only in combination with work devices that a tractor becomes a working machine. These devices may be attached, hitched, or mounted to the tractor. For rear device attachment, a three-bar linkage combined with a power lift has proved to be useful for the following reasons:

- Statically defined rigid connection between tractor and device
- Adjustable device due to the power lift
- The three-bar linkage can be adjusted to different devices by control devices
- The possibility of lateral mobility restriction or height and lateral mobility restriction

The three-bar linkage consists of two lower links and one upper link. The three-bar linkage's individual part dimensions are standardized into four categories (I–IV) for different power classes [14.31].

For a long time, rear hitches have been equipped with a control device that automatically lifts and lowers the power lift depending on a control value. By means of this power lift control, the driver is relieved of a task and working efficiency is increased by reducing the slip between the tyre and the soil as well as by raising the drive force. Possible control values are the device position relative to the tractor (position control), the traction force acting between the tractor and the device (traction force control), the device position related to the soil (depth control), and a mixture of traction force and position control (mixed control).

When using slip control, the actual driving speed is measured by a radar sensor. The theoretical speed is defined by a wheel-speed sensor. The electronics determines the speed difference and the slip. If the speed difference is smaller than a fixed limit, the electronic lifting control only works with traction force control. If the speed difference exceeds this limit, the traction force is changed by working depth modification so as to reduce the working depth and the slip. Another function of the electronic lifting control is active vibration removal, which involves balancing of the device's vibrations when passing on roads by means of automatic hydraulic steering in the opposite direction. This helps to improve driving comfort and steering conditions when driving with heavy devices attached to the tractor. The force-measuring bolts of the lower links serve to measure the force signal of the lifted device. The dynamic part emerged by vibrations is used as an actual value for lifting device control. In a restricted control area, for damping the lifting device is lifted or lowered slightly together with the attached device.

Control System and Electronics

The task of electronics is to control a tractor's components [14.32]:

- Motor
- Engine
- Hydraulics
- Gears
- Devices
- Data collection and storage
- Diagnosis

Modern tractors mainly use microcontrollers [14.26] (Fig. 14.31), equipped with CAN interfaces for communication. This yields new possibilities of diagnosis and overall optimization of the tractor's system. Due to the fact that tractors are applied with very different devices attached, which also using electronic control systems, it is important to make them communicate by a uniform interface, e.g., according to the ISO 11783 standard [14.33].

14.3 Machinery for Concrete Works

14.3.1 Concrete Mixing Plants

Since they must be well suited to their intended use and to construction needs, concrete mixing plants form a class of machinery with highly diverse designs. The diversity of concrete mixing plants' design features is connected with their capacity (10–250 m^3/h), their compatibility with the means of receiving the produced concrete mix, the production process control automation, the need for printing certificates for the sold concrete mix, the required assembly area, and the climatic conditions in which they are operated.

Concrete mixing plants can be classified as follows [14.34]:

- Concrete mixing plants (equipped with a mixer) producing ready-made concrete mix and concrete mix batching plants for proportioning and feeding

concrete mix constituents into truck mixers. Both kinds of plants can be built in vertical (tower concrete mixing plants) or horizontal configuration. Concrete mix batching plants may feed, depending on the time needed for transporting concrete mix to the destination, only dry constituents (aggregate and cement) or aggregate, cement, and water into truck concrete mixers.

- Stationary, transferable, and mobile concrete mixing plants. The kind of concrete mixing plant is determined by the operating costs, dependent on the length of time of operation at a given location.
- Continuous and batch production concrete mixing plants. Batch production concrete mixing plants are used on sites where the produced concrete

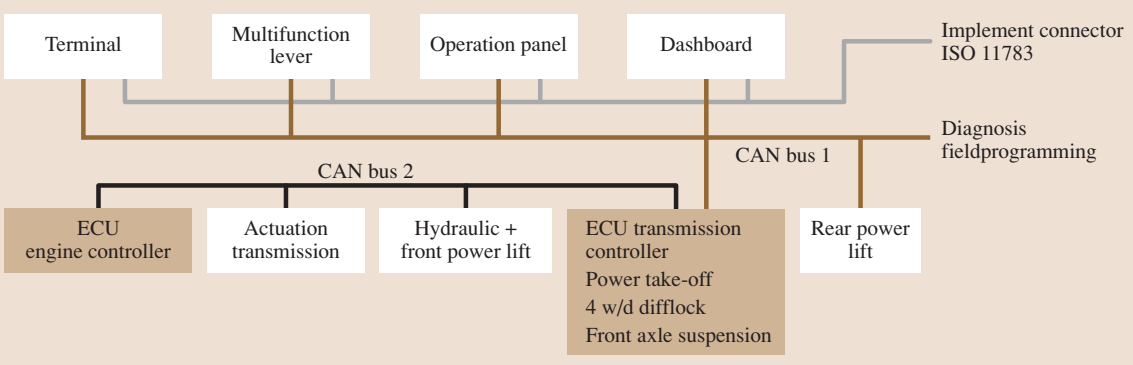


Fig. 14.31 Structure of a tractor control system

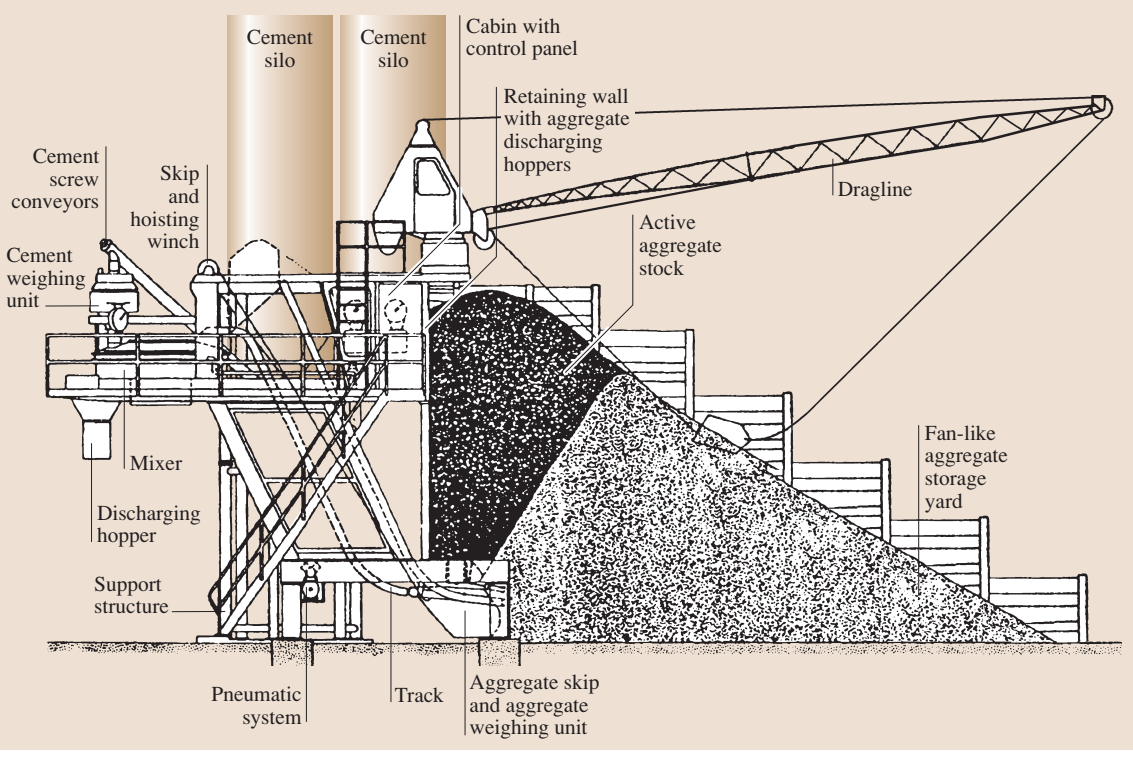


Fig. 14.32 Horizontal concrete mixing plant

mix's composition often needs to be changed, while continuous production mixing plants are used on construction sites which require large quantities of homogenous concrete, such as dams or airfields. Schematics of two basic types of batch production concrete mixing plants – horizontal and vertical (tower) – are shown in Figs. 14.32 and 14.33.

The concrete mixing plant's working cycle begins with the filling of the proportioners with aggregate, cement, water (if the latter is dosed by a weighing unit), and admixtures, which are then fed in the appropriate order into the mixer.

If flow-type water dosage units are used, water is fed into the mixer after the dry constituents have been charged into it. When the mixing process ends, the contents of the mixer is discharged into the recipient. Modern concrete mixing plants perform 50–60 working cycles per hour.

In order to achieve the desired technological capacities, the pressure of the water with which the concrete mix plant is supplied should be 0.4–0.6 MPa. The active aggregate stock (Fig. 14.32) represents the free-falling amount of aggregate which can be fed into the skip without using the dragline. The concrete mixing plant's personnel usually consists of two operators (operating the concrete mixing plant and the dragline). Nowadays automatically controlled charging skip hoists are increasingly being employed, allowing the operating personnel to be reduced to one person.

Because of their technological and economic advantages, horizontal concrete mixing plants are most commonly used in the construction industry.

The other type, vertical concrete mixing plants (Fig. 14.33), is characterized by the location of aggregate storage bins above the mixer. In vertical concrete mixing plants aggregate is usually transported to the storage bins by belt or bucket conveyors.

Because of the considerable height of vertical concrete mixing plants and the weight of their supporting structures they are constructed as stationary facilities within prefabricated concrete plants and as readymix concrete mixing plants for permanent locations. Their advantages include the ease of fitting them into concrete products manufacturing plants and adapting for concrete mix loading into overhead trolleys as well as easier adaptation to operation at low ambient temperatures. As a rule they are built for year-round operation.

Mainly electric motors are used in the power transmission systems of concrete mixing plants.

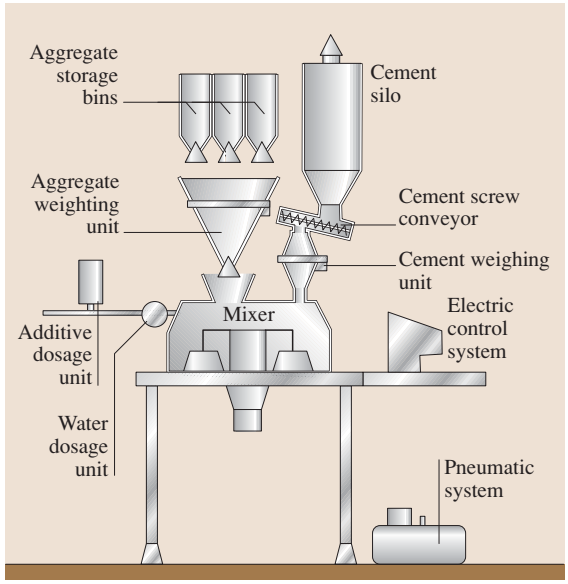


Fig. 14.33 Schematic of vertical concrete mixing plant

Only the closing of the devices batching concrete mix components is effected by pneumatic actuators because of the high speed of motion required.

Since it is capable of generating great forces, a hydraulic drive is usually used to open and close the mixer's discharging gate.

The most popular type of concrete mixing plant appears to be the transferable horizontal concrete mixing plant, which is characterized by a simple design and that lends itself to quick assembly on the construction site. The components of horizontal concrete mixing plants are transported to the site on generally available means of road transport. Concrete mixing plants specially designed for quick relocation are referred to as mobile. They are usually mounted on specially designed trailers.

Concrete mixing plants are usually built according to the needs of the individual user, who can choose the kind of mix, the height at which the concrete mix is discharged, the kind of aggregate storage facility (Fig. 14.34), the number of aggregate fractions and grades of cement, the batchers' weighing systems, the use of admixture and additive proportioners, the automatic make-up water dosage unit, the type of control system, the operator's cabin, adaptation to operation at low temperatures, and environmental compatibility.

Depending on the size of the concrete mixing plant, users can choose from simple concrete mix batcher weighing units with a spring head with an accuracy of ±0.5% or electronic ones based on strain gauges whose

accuracy, according to some manufacturer's specifications, is as high as ±0.1%.

An additional advantage of weighing units based on strain gauges is that the signals can be processed directly, whereas in older systems the weighing head indicator rotation had to be converted into an electric signal by means of coupled voltmeters or reed relays mounted on the head's dial.

The use of make-up water dosage devices which take aggregate moisture into account is essential when high concrete homogeneity is required.

Automatic make-up water dosing devices which take aggregate moisture into account can be divided into two groups:

- Devices determining the water content in aggregate (chiefly sand) and adjusting on this basis the amount of water fed into the mixer. The water content is determined by gauges, built in at the outlet from the sand storage bins, measuring electric conductivity, microwave flow, permittivity or the moderation of fast neutrons.
- Devices dosing water to the concrete mix on the basis of concrete mix consistency determinations. The consistency is determined by measuring the electrical resistance of the concrete mix. Gauges built into the mixer's bottom or electrodes mounted similarly as mixing blades are used for this purpose.

State-of-the-art systems for controlling the production process in the concrete mixing plant are available to order.

The application of personal computers (PCs) and programmable logic controllers (PLCs) represents a revolution in concrete mixing plant control systems, making it possible to use any number of recipes, produce batches constituting any percentage of the nominal batch of concrete, choose the number of batches, transfer data remotely to program recipes, print certificates for the manufactured concrete mixes, and accurately control concrete mix inventories and sales.

Contemporary concrete mixing plants must meet environmental requirements, firstly air protection requirements. This is ensured by sealing off (with closing flaps and dust filters) the mixers as they are filled with dry components and using efficient air–cement mixture filters in the cement silos. Also the concrete mix residues from washing the mixer must be utilized.

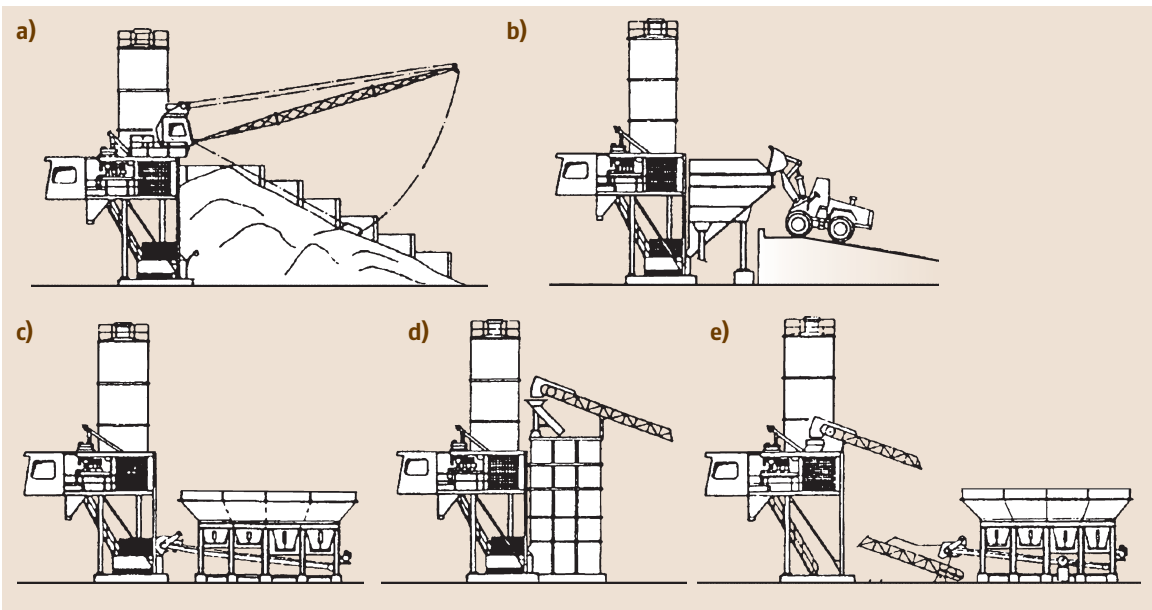


Fig. 14.34a–e Different aggregate storage solutions **(a)** Fan-like storage yard. **(b)** Fan-like storage bin with loading ramp for loader. **(c)** Row aggregate storage bins situated by concrete mixing plant, and belt conveyor. **(d)** Fan-like storage bin with aggregate distributor and belt conveyor. **(e)** Row aggregate bins with belt conveyor weighing and feeding aggregate into container located above mixer

14.3.2 Concrete Mixers

According to the definition given in [14.34], the concrete mixer is a machine designed for the production of concrete mix by mixing measured (by mass or volume) proportions of water, cement, aggregate, and chemical additives (if used) within a certain time limit. Depending on its design, the concrete mixer consists of a basic unit, called a mixer, and ancillary units such as: a wheeled supporting frame or a support structure, a charging skip (possibly with a weighing device) for transporting concrete mix components, a water dosage unit, a mechanical shovel, and a control box.

Concrete mixers form the largest group of construction machines. One can find them on nearly each construction site. They are highly diverse as regards size, which is defined in terms of dry component capacity or the volume of ready concrete obtained from one batch [14.35], and design.

There are concrete mixers with a dry components capacity of 50–12 000 l. Small concrete mixers with a capacity of 50–250 l usually work as single machines with manual transport of concrete mix components to the mixer. In order to use their potential rationally, concrete mixers with a capacity above 375 l should work in conjunction with mechanized concrete mix components transport, i. e., they should be incorporated into a concrete mixing plant.

Depending on the way they operate, concrete mixers can be classified as follows:

- Freefall (gravity) concrete mixers versus compulsory concrete mixers
- Batch- versus continuous-type concrete mixers.

Freefall mixers find application in monolithic construction for production concrete mixes with consistency ranging from liquid to plastic. They are usually used for the construction of residential buildings and livestock and public utility structures. Depending on their discharging method, three types of freefall mixers can be identified: tipping-drum concrete mixers, reversing-drum concrete mixers, and discharging chute concrete mixers. Tipping-drum concrete mixers with a dry components capacity of 50–250 dm^3, used for the production of concretes and mortars in family housing, form the most numerous group among gravity mixers (Fig. 14.35). Similarly as for electrical appliances, they are distributed by chain stores. The main hazard associated with their use is the possibility of electric shock. This hazard can be eliminated through the use of an integrated double-insulated motor switch unit.

To facilitate their transport on public roads concrete mixers can be fitted with towbars with a hook coupling, pneumatic tyres, a brake, and lights.

For more thorough washing after work the mixing drum is closed with a special cover so it can be rotated horizontally.

Tipping-drum freefall concrete mixers have a capacity of 350–1000 dm^3, whereas reversing-drum concrete mixers typically have a capacity of 350–1000 dm^3. Discharging chute concrete mixers are being phased out and replaced by mixers of the above two types. Their design used to be advantageous when combustion engines with one sense of rotation were employed to drive them whereby reversing gears did not have to be used. As electric drives have been widely introduced and the production of concrete mix could be automated, their design became obsolete.

Current development of freefall concrete mixers is directed towards improving their operational safety conditions, reducing noise emission, and facilitating their transport on public roads and maintenance.

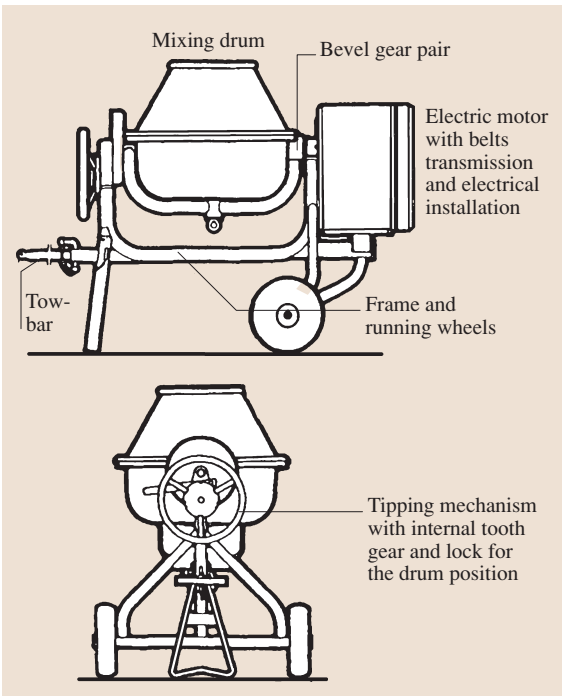


Fig. 14.35 Small-sized tipping-drum gravity concrete mixer equipped with traveling wheels

Compulsory concrete mixers are used for the production of all types of concrete mixes and, from the concrete preparation technology point of view, their application is unlimited.

Compulsory concrete mixers are divided into two classes: pan-type and paddle-type (trough-type) concrete mixers.

In pan-type concrete mixers, mixing is effected by the rotation of a set of agitators inside a pan with a vertical axle, whereas in paddle concrete mixers the set of agitators revolves in a trough with a horizontal axle.

Within the class of pan-type concrete mixers six types of machines can be identified: turbo concrete mixers, planetary concrete mixers, turbo planetary concrete mixers, countercurrent operation concrete mixers, concurrent operation concrete mixers, and concrete mixers with a high-speed stirrer. In the present-day construction industry and the concrete prefabrication industry the first three types of concrete mixers, i. e., turbo concrete mixers, planetary concrete mixers, and turbo planetary concrete mixers, are the most commonly used. The operation of these machines' agitators is illustrated in Fig. 14.36.

A planetary concrete mixer with a capacity of 1000 dm^3 is shown in Fig. 14.37.

As already mentioned, turbo concrete mixers, planetary concrete mixers, turbo planetary concrete mixers, and one- and two-agitator paddle mixers are the most popular types amongst the numerous group of compulsory concrete mixers. No clear advantage in terms of the quality of the produced concrete mixes of a single concrete mixer type over the other types is observed for ordinary concretes [14.35]. As regards special concretes (e.g., with a very low water/cement (W/C) ratio, extreme consistencies or made using non-mineral aggregates) the application of a particular type of concrete mixer should be agreed on between the purchaser and the manufacturer. When purchasing a concrete mixer one should also take into consideration the particular features of each concrete mixer type, such as:

- The possibility of using large-size coarse aggregates: concrete mixers with elastically suspended mixing blade arms are better suited for this purpose than those with rigid suspension of the mixing blade arms.
- The power demands of the mixing process: one can assume that the power demands for the mixing process in a planetary concrete mixer will be lower (by about 25%) than that of a turbo concrete mixer.
- Environmental compatibility: the design of the upper cover of the mixer should protect the environment against cement dust emission and concrete mix splashing, and the seal of the discharge gate should prevent concrete mix leakage.

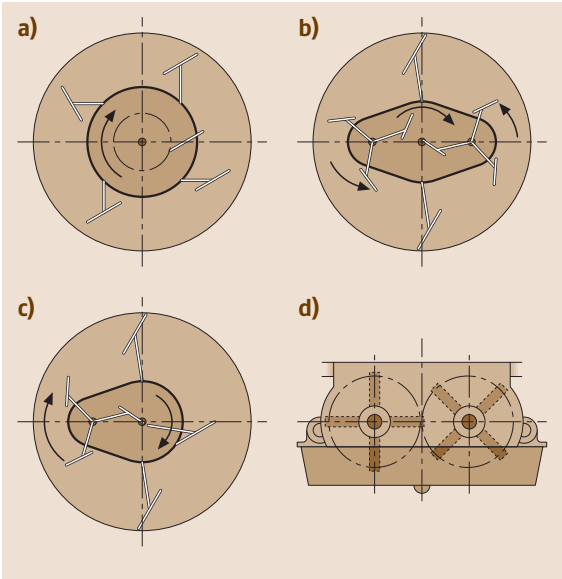


Fig. 14.36a–d Schematics illustrating operation of selected types of pan-type concrete mixers and two-agitator paddle concrete mixer

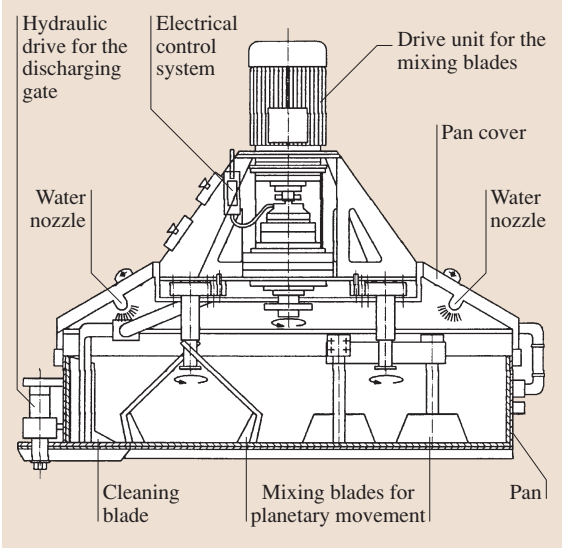


Fig. 14.37 Planetary mixer with a capacity of 1000 dm^3

- Protection against agitator drive system overload: the use of hydraulic torque converters prevents drive system overloads and ensures a smooth start of the rotation of the agitators as the mixer is filled; also elastic suspension of the mixing blade arms protects the drive against overloading and deformation of the mixing blade arms.
- Quick distribution of make-up water in the mass of dry components to reduce mixing time: several suitably arranged water nozzles ensure that the concrete mix quickly becomes homogenous.
- The concrete mixer's dimensions should be suitable for transport on public roads.
- The use of abrasion-resistant materials for the blades and the linings.

14.3.3 Truck Concrete Mixers

The truck mixer (Fig. 14.38) is designed for producing homogenous concrete mix and transporting it over long distances. It consists of a pear-shaped drum (usually a freefall, reversing one) inclined at an angle of 15° and a self-propelled chassis or a trailer. Its accessories include: a water tank, a water dosage unit, a charging hopper, and discharging chutes. The drum is supported by a cylindrical pin on the drive side and by two rollers mating with a rigid ring fastened on the drum. As concrete mix components are loaded into the drum and mixed, the drum revolves in one direction. During discharging the drum revolves in the opposite direction.

It is required that the mixing drum can rotate at several rotational speeds: at the highest speed while being filled or emptied (e.g., into a concrete pump), at a lower speed during mixing, and at the lowest speed during travel. For this reason a hydraulic drive is most commonly used for rotating the drum. This drive enables stepless change of the drum's revolutions in a range of 0–14 rpm in both directions. This ensures the most suitable mixing drum revolution rate for concrete mix loading, transport, and unloading, adjusted to concrete mix placement on the construction site. The truck mixer's hydraulic drive powered by the vehicle's engine is shown schematically in Fig. 14.39. It consists of a combustion engine, a hydraulic pump with an adjustable rate of delivery, a hydraulic engine, and a planetary gear to drive the mixing drum. The prime mover of the mixing drum in truck mixers can be an independent combustion engine or the engine that propels the chassis.

The factor limiting the distance over which concrete mix can be transported by truck concrete mixers is the setting time, determined mainly by the ambient temperature and the temperature of the concrete mix. If the expected transporting time is longer than the setting

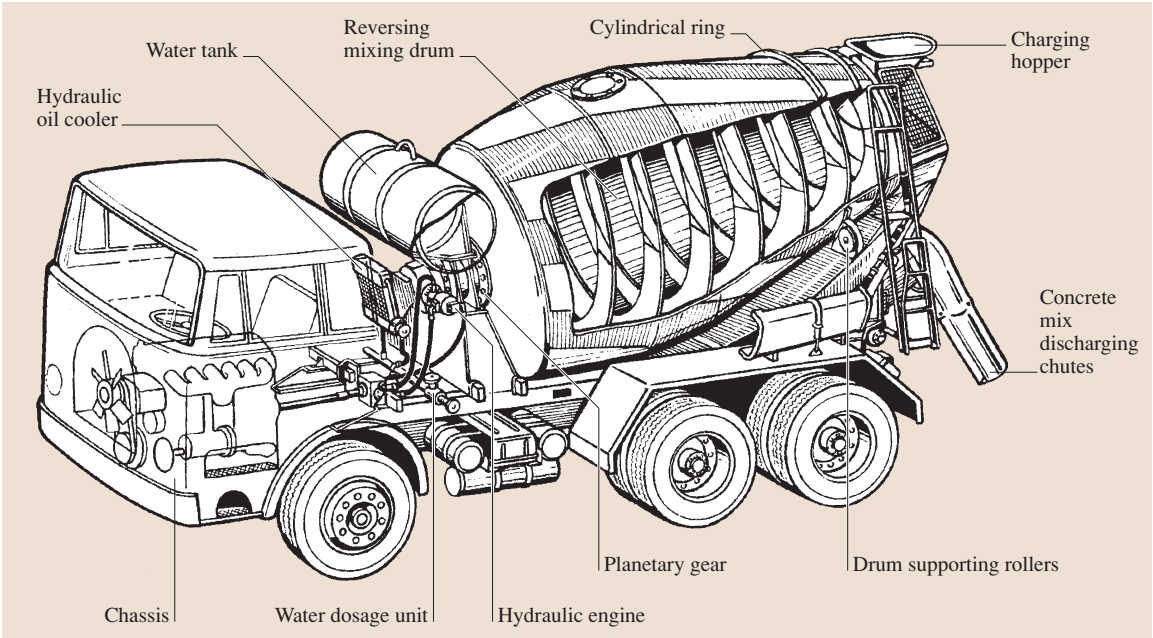


Fig. 14.38 Truck mixer with drum driven by the truck's engine

time, dry components are fed into the concrete mixer and, 35 min before discharge, water is added and mixing is started.

While the truck concrete mixer filled with both concrete mix and dry components is traveling the contents is mixed at a low rate (about 3 rpm) in order to prevent concrete mix segregation or cement setting in contact with moist aggregate.

The following trends in the development of truck concrete mixers can be observed:

- The use of mixing drums with an ever greater capacity – 9, 10, and 12 m^3 – aimed at reducing transport costs and making it easier to maintain continuity of concreting when erecting large concrete structures. In order to increase the load capacity of concrete mixers they are fitted out with an additional rotary axle at the back of the truck, which is raised during discharging and while driving without a load.
- The combination of a truck concrete mixer and a concrete pump with a distributing boom or a belt conveyor for direct placement of the concrete mix on small construction sites;
- The adaptation of the truck concrete mixer for concrete mix transport at low temperatures (as low as −60 °C). In this case, the mixing drum is made of two shells (an outer shell and an inner shell) and hot exhaust gas is fed into the space between the shells.

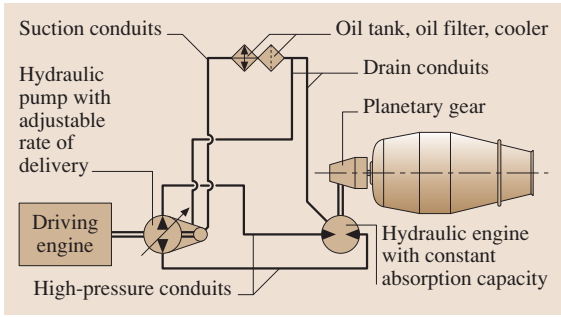


Fig. 14.39 Schematic of a truck mixer drum's hydraulic drive powered by the truck's engine

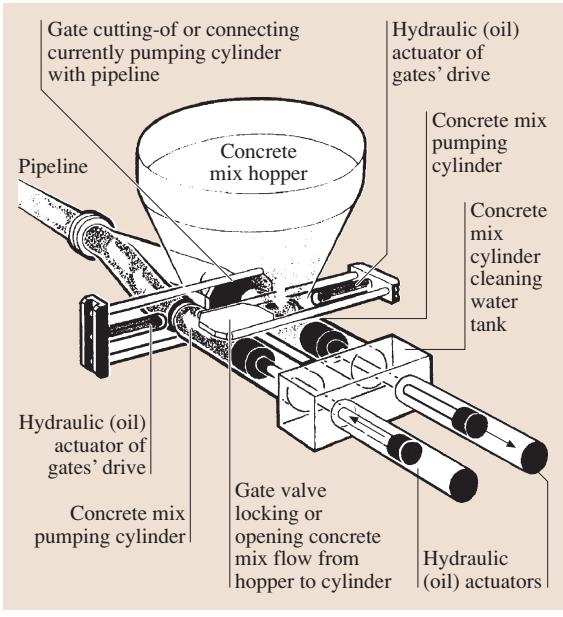


Fig. 14.40 Principle of operation of a concrete pump with control valves in the form of flat gates with connecting ports

14.3.4 Concrete Pumps

Concrete pumps have become the dominant means of concrete mix conveyance on the construction site. They have entirely supplanted pneumatic feeders, which deliver concrete mix in batches. The rapid development of concrete pumps started once hydraulic drives were incorporated into their design, supplanting crank power drives.

Another breakthrough in the development of concrete pumps was the use of distributing booms, which greatly facilitated the placement of of concrete in the work area by delivering concrete mix directly to the placement area and distributing it there.

Currently the most common pump design is a pump with two parallel, alternately operating concrete mix cylinders whose pistons are driven by hydraulic (oil) actuators connected to them in series (Fig. 14.40).

The reciprocating motion of the concrete mix pistons is controlled by means of two noncontact limit switches located in the cleaning water tank. Switches of the same type are also used to control the motion of the cutoff valves.

An important distinguishing feature of particular piston-type concrete pump designs is the system of valves that controls the flow of concrete mix from the hopper to the cylinders and from the cylinders to the conveying pipe.

Besides valves in the form of flat gates, other valve systems, such as the conveying pipe's swing segment (C- and S-valves), plug valves, etc., are employed. Attempting to bypass patented designs individual manufacturers of concrete pumps have developed new valve system designs for piston-type concrete pumps [14.36].

A further major advance in the development of systems for controlling the flow of concrete mix in concrete pumps was the use of C- and S-valves. Their operation consists in the alternate connection of the conveying pipe's swing segment fully (S-valves) or partially (C-valves – the elephant-type system) submerged in the concrete mix hopper while the cylinders are in the pumping phase.

In order to minimize wear, concrete mix pumping cylinders are characterized by large diameters, long length, and low rate of displacement. These parameters, in conjunction with reduced speed of the piston as it approaches the dead center and fast switching of the control valves, reduce pumping fluctuations so that the flow of concrete mix is almost continuous.

Pumps are selected depending on the type of structure to be erected. The extreme examples are high-capacity pumps with a short delivery distance versus low-capacity pumps and a long conveying distance. Depending on the application needs, pumps with concrete mix cylinders with a diameter of 100–280 mm and a forcing pressure of 8–26 MPa are used. Pumping capacities range from 20 to 200 m^3/h.

For some construction projects special pumps are built. In the technical literature one can find information about concrete mix pumped to an elevation of 530 m and over a distance of 4000 m (the construction of the Schaeftlarn tunnel). Such pumping distances are achieved when specially designed pumpable concrete mixes are pumped. The design features of these mixes include: the consistency, the cement content, the additives content, and the shape and grading of the aggregates. Another crucial factor is the ambient temperature. A high ambient temperature accelerates concrete mix setting and limits the pumping distance.

Concrete pumps (Fig. 14.41) are usually manufactured as self-propelled machines on chassis or trailers to be towed by a tractor. Concrete pumps mounted on self-propelled chassis are usually powered by the vehicle's engines while those mounted on trailers are driven by a separate engine.

Stationary pumps are fitted with skids or driving axles, which are dismantled on the construction site. They are driven by diesel engines or electric motors.

The choice of a concrete pump (self-propelled, mounted on a trailer or stationary) is determined by economic factors and the character of the construction project. One should take into account that the ratio of the concrete mix placing time to the formwork and to the reinforcement time is relatively low, roughly 1:5.5.

Self-propelled concrete pumps with a distributing boom (Fig. 14.42) ensure very efficient concrete mix transport and distribution.

The distributing boom consists of two, three, four or five segments with a delivery pipe with a rubber hose attached to it. The reach of pumps with a distributing boom is 17–65 m in the horizontal plane. An exemplary nomogram of a 19 m long pump boom's horizontal and vertical reach is shown in Fig. 14.43.

Booms require appropriate chassis and supports. The angle of rotation of the boom around the vertical axis is limited to 270° because the hydraulic conduits feeding the boom folding cylinders are located next to the concrete mix delivery pipe's articulated joint. In order to adapt them better to the characteristics of construction sites, the distributing masts are built as foldable from top or bottom. The folding and slewing of the mast is effected by push-button control. In the case of five-segment masts, folding is effected by a control system following a program.

The approximate commercial specifications of mass-produced self-propelled concrete pumps with a distributing boom are as follows:

Reach in horizontal plane:	12–58 m
Reach in vertical plane:	16–62 m

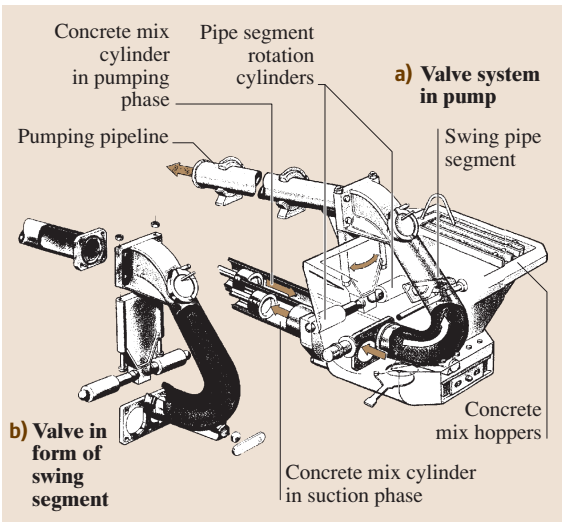


Fig. 14.41a,b Principle of operation of a concrete pump with a valve system in the form of the conveying pipe's swing segment connecting conveying pipe with cylinders during the pumping phase (C-valves – elephant-type system). **(a)** valve system in pump; **(b)** valve in the form of a swing segment

Slewing angle:	360°
Conveying pipe diameter:	100–125 m
Length of rubber hose for	
Distributing concrete mix:	3.5–5 m
Charging hopper capacity:	350–500 dm^3
Rated pumping capacity:	36–150 m^3/h
Concrete pumping pressure:	45–130 bar
Concrete mix pumping	
Cylinder diameter:	160–230 mm
Piston stroke:	1000–2100 mm
Number of boom segments:	2–5

Self-propelled concrete pumps are also made in versions adapted for attaching a pipeline made from steel pipes for conveying concrete mix over considerable distances.

Besides piston-type concrete pumps also rotary-type pumps, based on the principle of the peristaltic pump, are manufactured (Fig. 14.44).

In the rotary-type concrete pump concrete mix is pumped as a result of squeezing it out of a reinforced rubber hose by two rollers attached to the rotor. The hose recovers its circular cross-section owing to elastic restoration or negative pressure inside the casing (vacuum restore). The pumping pressure in rotary-type pumps amounts to about 3 MPa, allowing concrete mix to be delivered over a distance of about 200 m in the horizontal plane and to an elevation of about 80 m.

The design of the rotary-type concrete pump is simple but the conveying hose needs to be replaced quite often.

Concrete pumps' conveying pipelines are made from 0.5, 2, and 3 m long steel pipes usually 125 mm in diameter, but pipes 100, 150, and 180 mm in diameter are also used.

The pipeline attachments include pipe fittings with 90°, 120°, 135°, 150°, and 165° angles of flare. Individual pipes and pipe fittings are connected by clamping rings and rubber gaskets. The pipeline is tipped with a flexible conduit which facilitates the distribution of concrete mix in a work area.

After work, concrete mix residues are removed by a porous rubber ball forced through by a water jet or compressed air.

Besides the above self-propelled, trailer mounted, and stationary concrete pumps, there are also pumps mounted on tower cranes. These are used on construction sites with a large amount of reinforcement or difficult access for concrete mix placement by other methods, when large-diameter cylindrical tanks are to be built, and so on.

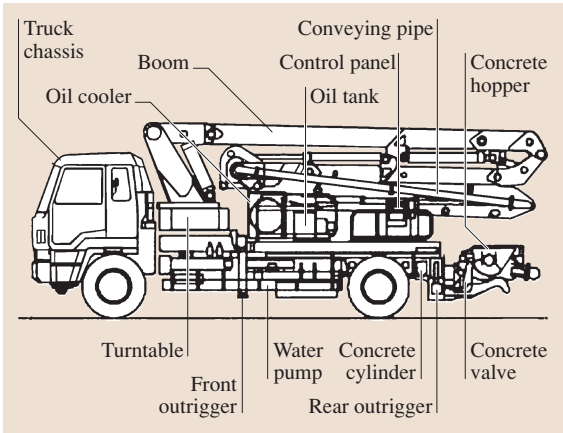


Fig. 14.42 Truck-mounted pump with distributing boom

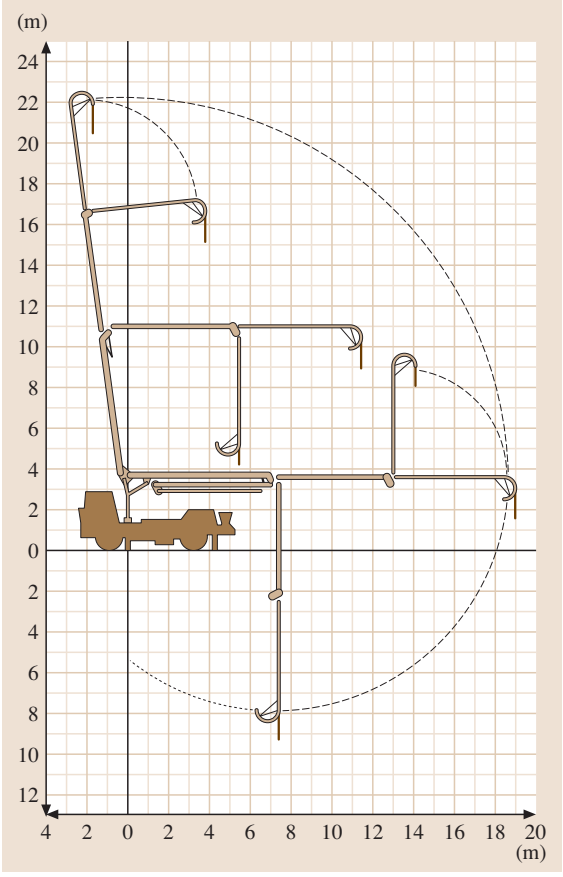


Fig. 14.43 Horizontal and vertical reach of pump with 19 m-long boom

14.3.5 Concrete Spraying Machines

Concrete spraying machines are used for spraying concretes and mortars onto structures. They are mainly used to coat the outer surface of the reinforcement in reinforced concrete structures, steel structures, tensioning cables, slope reinforcements, etc., to protect them against corrosion. Concrete mix can be sprayed onto concrete, rock, and steel bases, brick walls, and wooden formwork. Various kinds of cement concrete, including polymeric concretes and epoxy concretes, can be used for spraying. Sprayed concrete is characterized by good adhesion to the base and chemical resistance.

Besides being used for spraying concrete, concrete mixture sprayers can also be used for sand blasting and conveying concrete mix. In the literature on the subject one can find a third use for concrete sprayers, i. e., semi-wet spraying, which is particularly recommended when concrete mix is to be transported over long distances. Two main modes of operation of concrete sprayers are distinguished [14.37]:

- Dry spraying
- Wet spraying

The operation of a dry mixture sprayer consists of feeding dry components (cement and aggregate) into a charging hopper and pneumatically conveying them to a spraying nozzle, where water is added under a pressure of 0.4–0.6 MPa.

The advantage of the dry method is the possibility of spraying a layer of high-strength concrete owing to the low W/C ratio. A dry mixture sprayer designed for small-sized spraying works is shown in Fig. 14.45.

The machine's main assembly is a rotor mixing the material contained in the space where atmospheric pressure prevails with compressed air at an overpressure of 0.5–0.6 MPa. The principle of operation of the rotor is

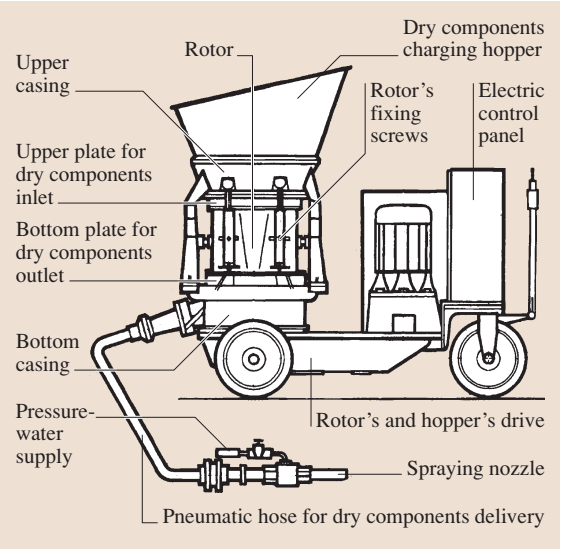


Fig. 14.45 Dry mixture sprayer for small-sized spraying works

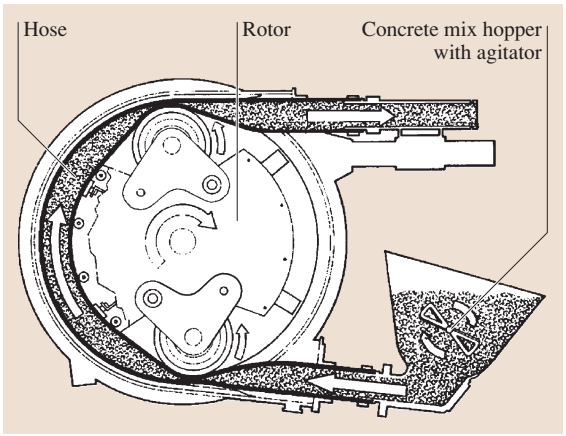


Fig. 14.44 Principle of operation of rotary type concrete pump

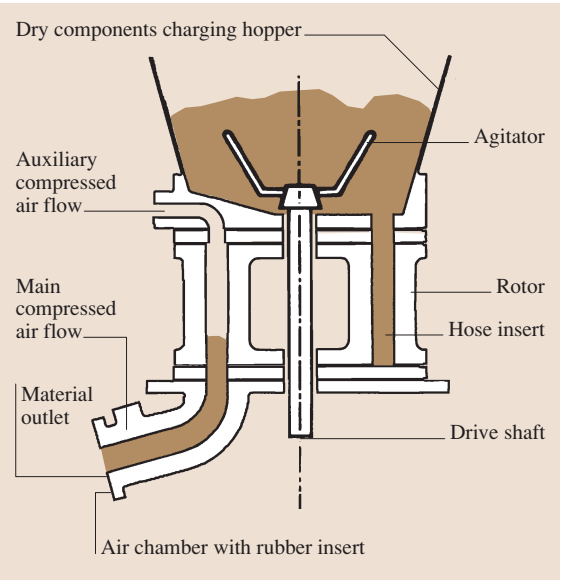


Fig. 14.46 Principle of rotor operation (example of rotor for dry spraying)

illustrated in Fig. 14.46. The aggregate and cement fed into the charging hopper under gravity fall into ports in the rotor and are forced by compressed air into the hose.

Dry sprayers' spraying capacity ranges from 0.2 to 11 m^3/h. Hoses 32, 38, 50, 60, and 65 mm in diameter are most commonly used.

It should be noted that concrete sprayers can also be used to transfer dry concrete mix components over considerable distances – up to 300 m in the horizontal plane and 40 m in the vertical plane – and to sandblast structures.

In wet sprayers ready-made (prepared in separate devices) concrete mix or mortar is conveyed to a spraying tip. In the group of dry sprayers two types of machines can be identified:

- Sprayers equipped with a rotor, in which concrete mix is delivered by compressed air (Fig. 14.47)
- Sprayers in which concrete mix is fed into a spraying tip by means of a concrete pump (Fig. 14.48)

In concrete-pump-type sprayers the concrete mix is conveyed by means of a pump and in the final stage also by compressed air, and is sprayed by a nozzle. Apart from nozzles with compressed air supply, impeller-type spraying units may be used as well.

In smaller devices the spraying nozzle is usually manually guided, whereas in large machines, mainly used in tunnel construction and mining operations, distributing booms wire-guided from portable control consoles are employed. Distributing booms can be mounted as self-contained machines mounted on wheeled or crawler carriers or form an integrated unit with a sprayer as shown in Fig. 14.48. A recent invention, besides nozzles spraying concrete mix by means of compressed air, is impeller-type spraying tips, where concrete mix is sprayed by a rotating impeller, using the centrifugal force directed by the latter assembly. According to the manufacturers, the impeller-type tip causes less dust and the concrete mix losses due to its bouncing off the wall are smaller.

Besides dry or wet mixture sprayers, there are also rotor-type sprayers in which dry and wet mixture spraying processes can be performed alternately after quick reconfiguration.

14.3.6 Internal Vibrators for Concrete

Internal vibrators for concrete are immersed in the concrete mix to transmit vibrations directly to it and thereby cause its compaction. As a result of this vibration the viscosity of the concrete mix decreases and its particles shift quickly relative to one another. The particular grains slide against one another and, due to settling under gravity, expel some of the air and water contained in the concrete mix, resulting in compaction and greater strength of the concrete after curing. When erecting concrete structures, in order to compact concrete mix effectively by means of immersion vibrators appropriate (in terms of frequency and amplitude) vibrators must be used.

The radius of action of an immersion vibrator depends on the exciting force and the frequency and ranges from 20 to 100 cm. Depending on the drive system, flexible drive immersion vibrators, built-in motor type electric immersion vibrators, pneumatic immersion vibrators, and hydraulic immersion vibrators are distinguished [14.38].

Flexible drive immersion vibrators are designed for driving a pendulum (Fig. 14.49) or eccentric vibration generator (Fig. 14.50). Pendulum-type immersion vibrators are driven by an electric motor or a combustion engine and a flexible shaft to which a vibration generator rolling on a raceway is attached.

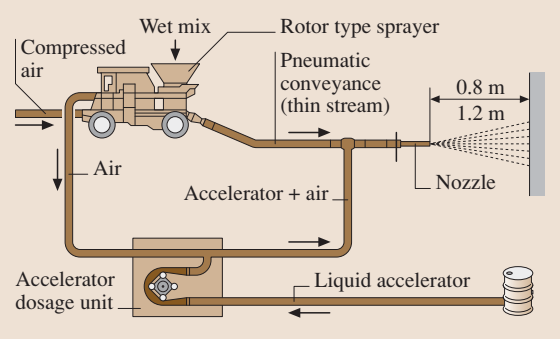


Fig. 14.47 Wet process with use of rotor-type sprayer

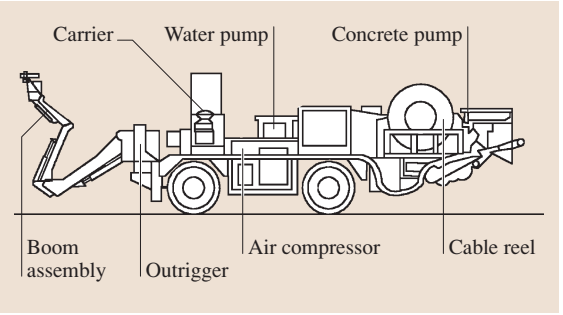


Fig. 14.48 Concrete-pump-type spraying machine (wheeled type)

The diameters of the vibration heads of the vibrators usually range from 25 to 70 mm and the frequencies generated are 300–200 Hz, respectively.

Modern pendulum-type immersion vibrators with a flexible shaft are driven by single-phase commutator motors; under load they can generate frequencies as high as 200 Hz (Fig. 14.50).

The diameters of vibration heads driven by commutator motors are usually 25–65 mm.

Built-in motor type electric immersion vibrators (Fig. 14.51) usually operate in conjunction with a voltage and frequency converter supplying a voltage of 42 V at 200 Hz. They can also be supplied from generating sets with an appropriate rated frequency. Because of the considerable permissible length of the power lead from the generator to the vibration head (about 15 m) these vibrators are suitable for compacting high elements. The vibrator's vibration heads are usually 35–85 mm in diameter and the vibration frequency is 200 Hz.

Pneumatic immersion vibrators are made with vibration heads that are 25–140 mm in diameter. They are characterized by simple design and high durability. These vibrators are used in places where, for safety reasons, it is inadvisable to use electric immersion vibrators or combustion engine immersion vibrators and in places with access to compressed air supply. Depending on their design, their frequency ranges from 160 to 300 Hz.

Hydraulic immersion vibrators are made with an eccentric vibration generator coupled with a hydraulic engine. They are used for compacting concrete mix when building large structures such as dams, bridges, and large foundations. The vibrators are supplied from special hydraulic feeders or from the hydraulic systems of earthmoving machines.

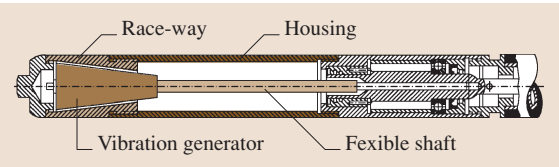


Fig. 14.49 Pendulum-type immersion vibrator

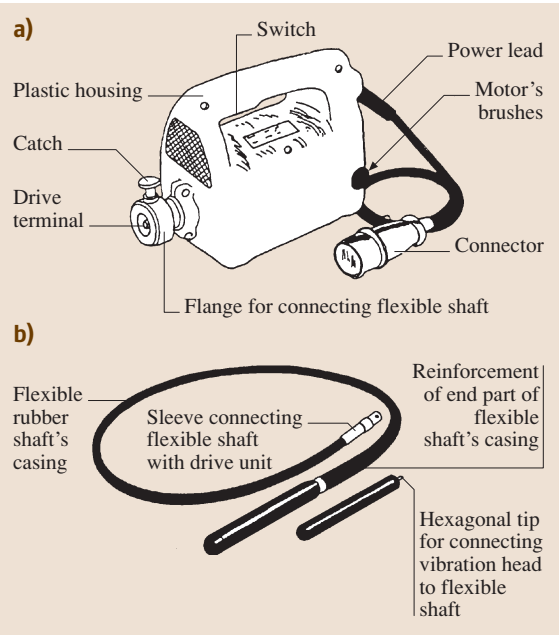


Fig. 14.50a,b Eccentric-type immersion vibrator driven by single-phase commutator motor and flexible shaft: **(a)** drive unit; **(b)** flexible shaft and vibration head

14.3.7 Vibrating Beams

Vibrating beams are used for leveling, compacting, and preliminary smoothing of the top surface of fresh concrete while building horizontally formed structures such as concrete road surfaces, airfields, storage yards, concrete floors in dwellings, and factory floors.

A vibrating beam consists of a rigid beam with a vibration generator mounted on it and flexible connectors for moving the beam. A vibrating beam with a vibration generator in the form of an attachable electric vibrator supplied with a voltage of 42 V is shown in Fig. 14.52.

In order to maintain the rectilinearity of the leveled surfaces and because of the considerable vertical gravity loading and exciting force loading, vibrating beams must have high vertical bending rigidity and relatively low deadweight so that they can be easily carried from one place to another on the construction site. These

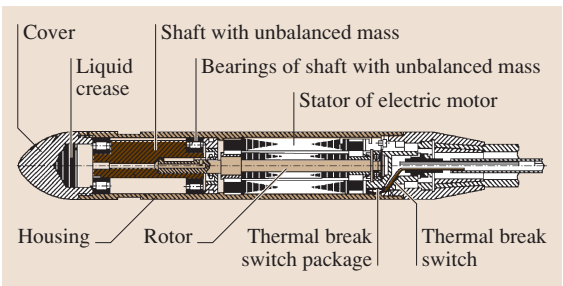


Fig. 14.51 Built-in motor-type electric immersion vibrator

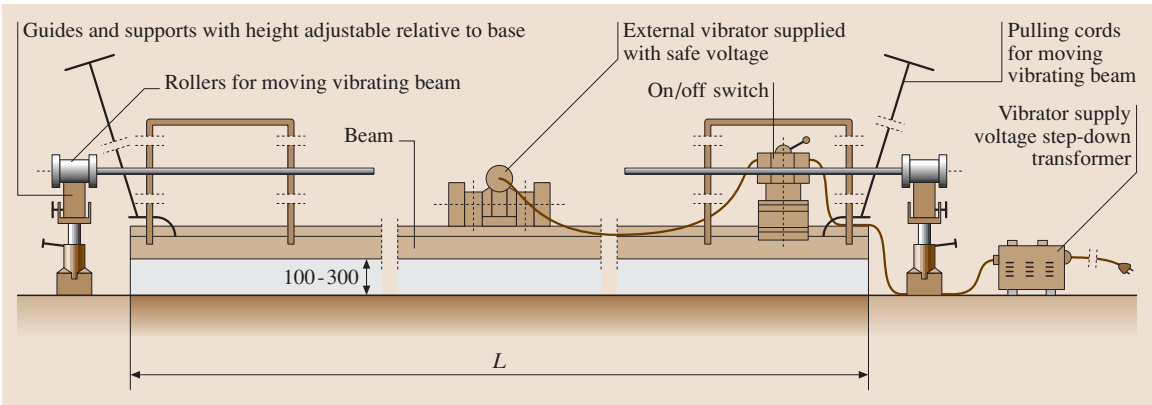


Fig. 14.52 Vibrating beam with vibration generator in the form of an external electric vibrator

requirements are met through the use of beams made from special multihole aluminum sections and by imparting negative deflection to the beams through built-in struts. The beams are used as single elements or double elements connected by lacings.

In order to reduce the harmful effect of vibration on the operators' hands, pulling cords or specially protected handles are used. Attachable vibrators with adjustable exciting force, driven by electric motors, combustion engines or compressed-air engines, are used as the vibration generators. Vibrating beams equipped with attachable electric vibrators are often adapted to a safe voltage supply to eliminate the electric shock hazard.

Vibrating beams of the type shown in Fig. 14.52 are manufactured 2.7–5.7 m long and can compact 10–30 cm thick bases.

The development of vibrating beams is directed towards increasing their length, obtaining uniform compaction of concrete mix along the whole length of the vibrating beam, and improving operational safety and transportability on the construction site. To a large extent the above requirements are met by the multipoint pneumatic vibrating beam shown in Fig. 14.53. The vibrating beam is 6.1 m long and consists of two end segments and one middle segment, joined together by bolts. The individual segments have a lattice structure triangular in cross section. Two stringers, one in the form of an angle and the other a T-bar, form the vibrating beam's base. The upper stringer is made from a pipe, which also serves as the compressed air conduit that supplies the vibrators. The vibrating beam is equipped with 16 pneumatic vibrators spaced at different intervals on both the lower stringers. The vibrators generate vertical vibrations that are transmitted to the concrete mix. The intensity of vibration is linked to the exciting force, the amplitude, and the frequency, and can be adjusted by controlling the supply air pressure. The upper stringer is joined together by turnbuckles which are also used to set the vibrating beam's deflection. The end segments are equipped with manually operated hoisting winches. The above vibrating beams are made up to 18 m long.

Vibrating beams can be equipped with different number of vibrators, depending on the length of the vibrating beam, the thickness of the compacted layer, and the desired surface smoothness. In practice, however, up to two vibrators driven by an electric motor or a combustion engine are used. In the case of reciprocating-motion pneumatic vibrators, this number is about 2–3 vibrators per running meter.

In recent years 1–3 m long smoothing vibrating beams have been introduced. These vibrating beams have low mass and are designed for smoothing elements made of semiliquid concrete mixes. The beam

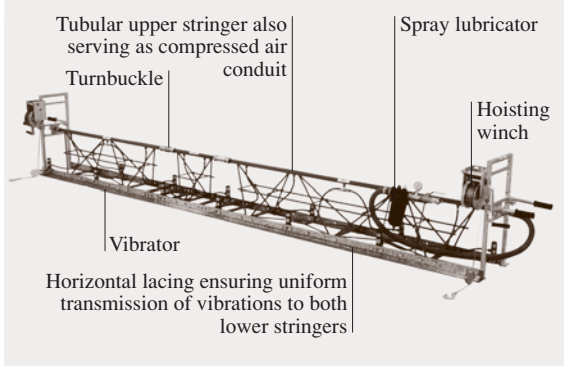


Fig. 14.53 Multipoint pneumatic vibrating beam

is equipped with an attachable electric or combustion engine vibrator and connected in an articulated way to a drawbar guided by one operator. In order to obtain high rigidity the vibrating beam's beam is made from properly formed (by bending) steel or aluminum sheet. Such beams are equipped with one vibrator, driven usually by a combustion engine or an electric motor.

14.3.8 Floating Machines for Concrete

The aim of floating is to obtain a high-quality concrete surface. Floating results in a level surface, a better compacted top surface, greater resistance to abrasion, corrosion, and the destructive effect of moisture, and a reduction in dusting. The floated surfaces can be painted or different kinds of flooring can be glued to them. Floating is performed after the base has gained some mechanical strength.

Rough and finishing floating are distinguished. In the case of older types of floating machines, rough floating was effected using a slowly rotating solid disk, while finishing floating was effected using blades set at an appropriate angle to the base. Now this practice has changed and the entire process is carried out using blades and changing the angle of their inclination, though solid disks are also used.

Single-disk floating machines (Fig. 14.54) are used for smaller concrete works. The working tools are blades or solid disks that are 600–1200 mm in diameter. They are driven by combustion engines or electric motors with a power of up to 8 kW. Combustion engine floating machines feature stepless control of the floating tools' rotational speed. Electric floating machines are equipped with two-speed electric motors whereby one can select the appropriate speed of rotation for the rough and finishing floating tools. Modern single-disk floating machines have the following features:

- A long handle, enabling access to floated surfaces with no need to walk on them (during transport of the floating machine the shaft is folded).
- The angle of inclination of the blades can be adjusted from the handle.
- The machine is equipped with a safety cutout switch (the so-called dead-man's grip), which automatically stops it once the operator's grip on the handle is released.
- Electric protection against switching the opposite direction of rotation.
- Road wheels for short-distance transport.
- The blades and the floating disks are made of high-quality materials to ensure long lifetime.

Two-disk floating machines (Fig. 14.55) are used for floating large surfaces.

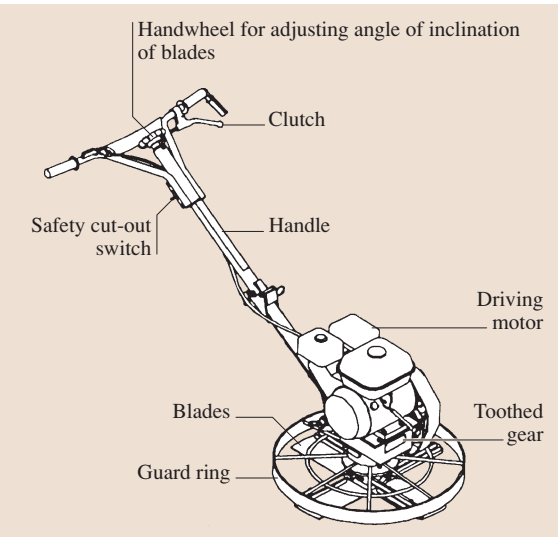


Fig. 14.54 Single-disk floating machine for concrete

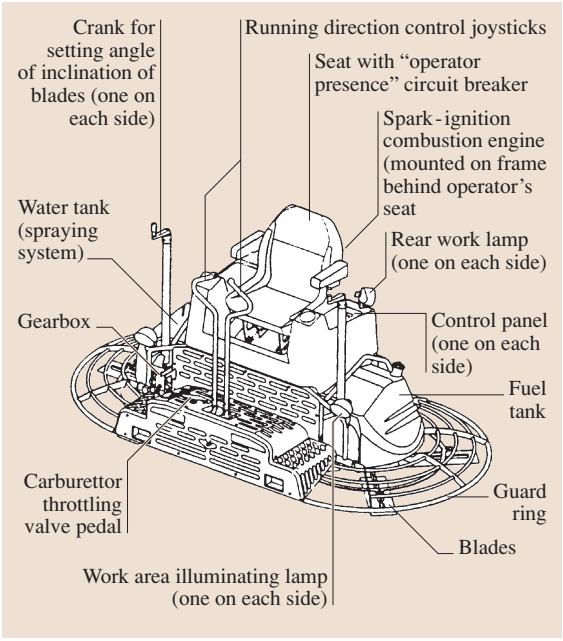


Fig. 14.55 Two-disk floating machine equipped with blades

For these machines the width of the floated strip in one pass ranges from 1700 to 2400 mm. The floating machine's control system consists of push-buttons, two joysticks for controlling the running direction, and two knobs for setting the angles of inclination of the blades.

The operator controls the floating machine while sitting in a centrally situated seat whose symmetry axis coincides with the machine's lateral axis. There are also tandem systems in which the symmetry axis of the operator's seat coincides with the machine's longitudinal axis.

In two-disk floating machines two floating system designs are used. In one of them the floating blades' outline circles may overlap while in the other there is a gap between the outline circles. The former design makes it possible to cover the whole floating width but no solid floating disks can be used. In the latter design floating disks can be used but the pass path needs to be corrected.

Depending on the floating width, machines of this type usually weigh 300–450 kg and the engine's maximum power rating is 24 kW. In the latest designs the mechanical systems are replaced by hydraulic systems, which ensure the smooth operation of the machine.

In two-disk floating machines the angle of inclination of the blades is adjusted by flexible-connector-type control systems. The control of the floating machine's running direction is based on the principle of differentiation of the friction forces acting in the particular quadrants of the outline circles, by changing the inclination of the blade ring. This principle applies to both one- and two-disk floating machines. Control is effected by means of the cranks, for each blade ring independently.

In order to ensure operational safety, safety cutout switches, usually in the form of a pedal pressed during operation by the operator's foot, are used in two-disk floating machines. If the operator falls out of the seat, the floating machine is automatically stopped.

14.3.9 Equipment for Vacuum Treatment of Concrete

Vacuum treatment is used to make high-quality concrete bases and floors. Its advantages include:

- Rapid increase in concrete strength in the initial period after placing concrete mix and applying the vacuum process
- A 15% increase in the final strength and the resulting cement savings
- Improved concrete features such as frost resistance, compression strength, imperviousness to water, and reduction in shrinkage deformation and floor dusting
- Reduction in the harmful effect of low temperatures on the curing of fresh concrete
- Quick execution of works

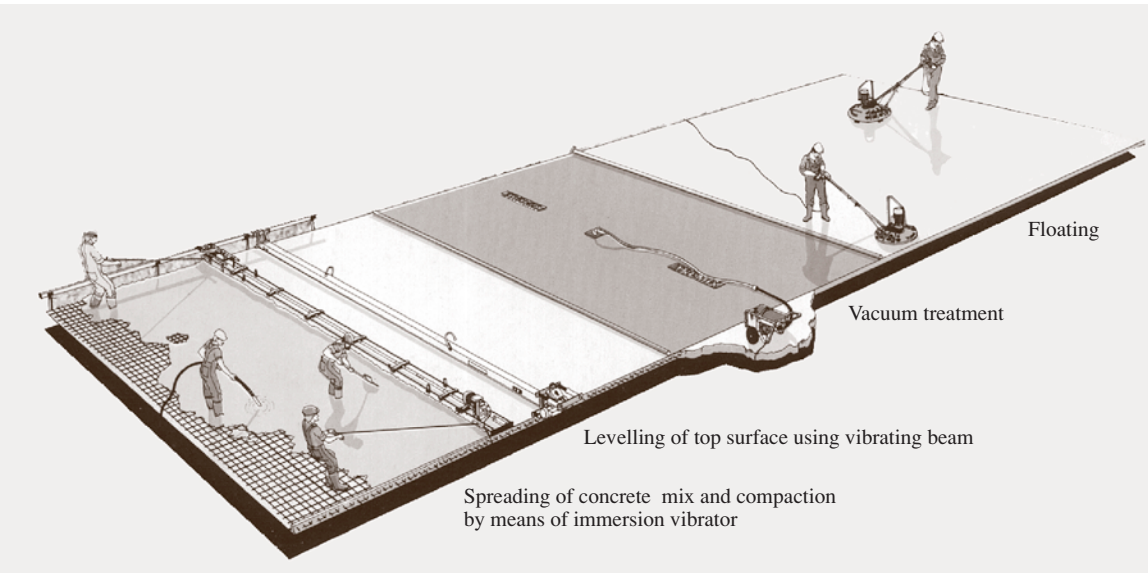


Fig. 14.56 Execution of concrete base with vacuum treatment use

- Ease of spreading of the concrete mix because of its semiliquid consistency

The use of vacuum processing of concrete is especially advantageous at low ambient temperatures (down to −5 °C) since the removal of excessive water and air bubbles to a large extent eliminates destructive processes.

The execution of concrete bases by the vacuum process can be divided into four operations (Fig. 14.56):

- Spreading of concrete and compaction by an immersion vibrator
- Compaction and leveling of the concrete mix's top surface by means of a vibrating beam pulled on guideways
- Covering with a suction mat and vacuum treatment
- Floating of the surfaces by means of rough and finishing floating machines

Roughly, the rate of vacuum treatment is 2 min/cm of base thickness. This means that a 10 cm thick base is treated for about 20 min. Floating is started when a boot impression in the concrete is about 3 mm deep.

The floating equipment includes immersion vibrators, vibrating beams, guideways with supports, expansion-joint inserts, vacuum unit with a suction mat, and floating machines.

The vacuum unit's main assembly (Fig. 14.57) is a vacuum pump with a driving motor. The pump usually has a sealing water-ring. The vacuum unit also includes: a vacuum tank, functioning as a settling tank for the sucked in impurities, connected to an atmospheric tank from which the air and water that has been sucked in is carried off, a wheeled frame, and an electric system.

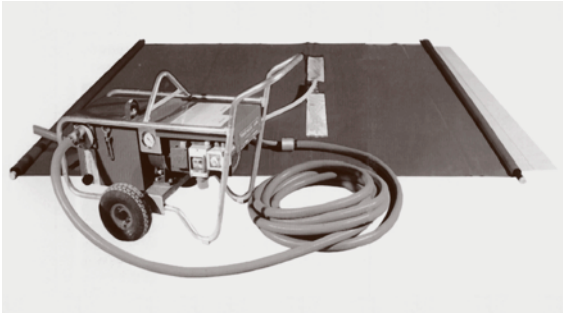


Fig. 14.57 Concrete vacuum treatment unit and suction mat

The development of vacuum units is directed towards reducing the mass and size of the vacuum unit and improving the mat. The currently used vacuum units made by leading manufacturers enable one-time sucking off of a concrete base 60 m^2 in area with a 4 kW driving motor and a machine mass of about 90 kg. The releaent pressure range for vacuum treatment of concrete is 75–95% vacuum (0.75–0.95 kg/cm^2 negative pressure).

Great advances have been made in the design of the suction mat. Older mat designs consisted of three layers, in the form of a filter cloth, a plastic flow mesh, and a tight cover, laid in turn on the vacuumed concrete mix. Current mats are manufactured as one integrated cover performing all three functions, i. e., ensuring the filtration and flow of the sucked off water and air and providing tight covering.

14.4 Site Lifts

14.4.1 Material and Equipment Lifts

Construction material lifts are intended for the vertical transport of building materials during the erection of new buildings and repairs. They may also be employed for the assembly of scaffolds and other construction site protection structures.

A classification of material and equipment lifts according to different criteria is shown in Table 14.1.

The most common type on construction sites are mast lifts with a cable or rack-and-pinion hoisting gear. Depending on the lifting height the lifts can be operated free-standing or anchored. The maximum lifting height of free-standing lifts depends on the stability of the supporting structure. The maximum lifting height of an unanchored lift does not usually exceed 12 m.

Mast Material and Equipment Lifts with Cable Hoisting Gear

The lifting capacity of lifts with a cable hoisting gear usually is below 600 kg.

Examples of lifts with a cable carriage hoisting drive are shown in Figs. 14.58, 14.60 and Table 14.2.

A lift with a capacity of 200 kg (Fig. 14.58) is made up of the following structural units:

Table 14.1 Classification of material lifts

Classification criterion	Lift definition
Platform drive design	Material and equipment lifts with cable hoisting gear
	Material and equipment lifts with rack-and-pinion hoisting gear
Platform track supporting structure	Material and equipment mast lifts
	Material and equipment shaft lifts
Base stability securement	Free-standing material and equipment lifts
	Anchored material and equipment lifts

- A base
- A multisectional mast
- A head with cable pulleys
- A drive unit-cable winch
- A carriage

The carriage moves along the mast, which is secured to the base. The mast is made from sections and has a segmental structure. It can be extended by adding more sections. The mast may be a free-standing structure or a structure anchored by means of special elements to the building's wall or window openings. The carriage is hoisted by a cable system. The winch's cable is guided to the mast's head, where there are cable pulleys through which the cable winds. The cable's end is secured to the carriage, which is equipped with rollers. The carriage suspended on the cable moves along the mast's guides. Accessories for transferring different materials are fixed to the carriage (Fig. 14.59). The carriage can slew around the mast by an angle of about 90°, which improves operating safety during the unloading of materials.

To facilitate the transport of the lift, its base is equipped with vehicle wheels.

Another 500 kg-capacity lift with a cable working platform raising drive is shown in Fig. 14.60.

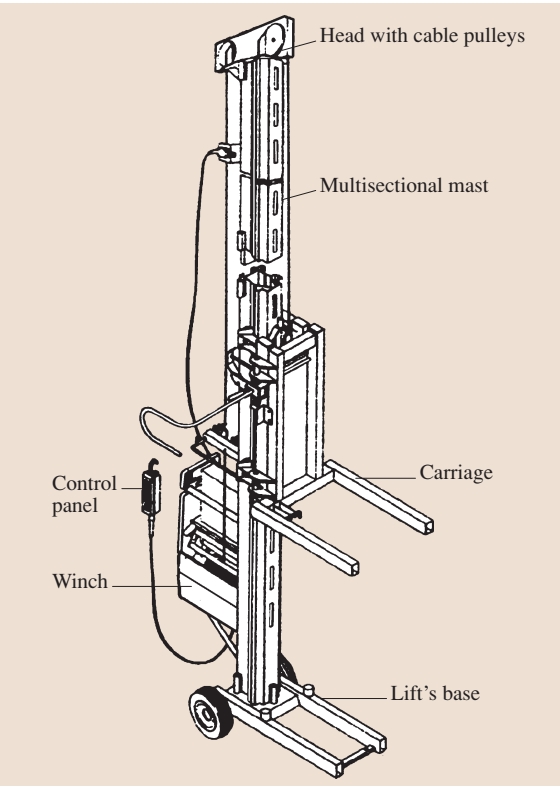


Fig. 14.58 Two-hundred-kilo-capacity lift with cable hosting drive

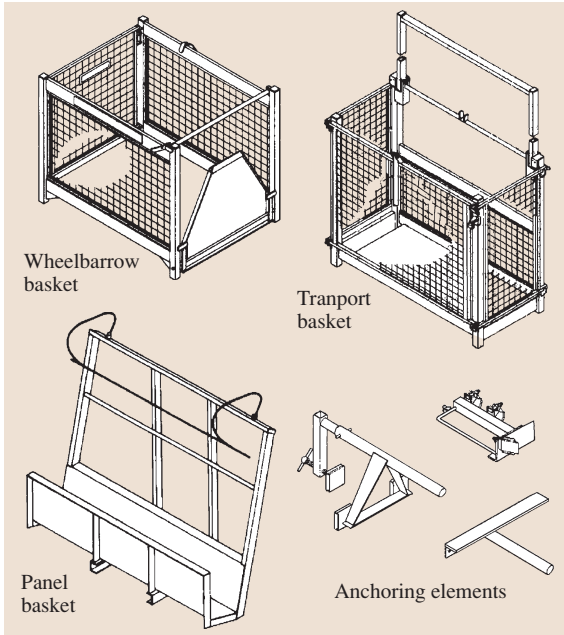


Fig. 14.59 Lift accessories

The lift consists of the following structural units (Fig. 14.60):

- A base
- A cable winch
- A transport basket
- A carriage
- A multisectional mast
- A head with cable pulleys

The basket can slew around the mast by 90°.

The lifts are self-assembled using the provided accessories.

Lifts with a cable hoisting drive are equipped with the following control and safety systems:

- Gripping devices activated when the limit speed is exceeded
- Limit switches making it possible to stop the lift at set stop levels
- A carriage (with the platform installed) upper position limit switch
- A supply-failure emergency carriage lowering system
- Stops in the lift's base

Ladder Lifts

Ladder lifts are cableways for transferring a load simultaneously in the vertical plane and in the horizontal plane. They are employed in the construction and repairs of residential buildings up to 30 m high and are used for roof and indoor work. Ladder lifts enable the transfer of building materials and the removal of waste materials. A typical ladder lift is shown in Fig. 14.61 and Table 14.3.

The ladder lift's main components are: a cable winch, a track, and a carriage.

A cable winch (Fig. 14.61) with a 230 V single-phase, squirrel-cage electric motor is the most popular drive unit used in ladder lifts. The winch includes the following assemblies: an electric motor, a roller gear, a cable drum with a cable, and an electrical system (contactors, relays, overload protections, 230 /24 V transformers).

Winches are mounted on frames attachable to a track segment. They are usually mounted between the truck's two lowest rungs. The drive unit is controlled via a portable control panel. The drive units employed in ladder lifts are very similar to the portable winches described in Sect. 14.6.2.

The main unit of ladder lifts is the track set at an angle for work (Fig. 14.61) It rests against the ground at the bottom and against a structural member (a window frame, the roof's edge) of the building at the top. When a ladder lift is to be used on a sloping roof, a track bend (Fig. 14.61) whose inclination can be adjusted is incorporated and then the track end segments (Fig. 14.61) that are to rest on the roof are attached. The track's section from the base to the building's structural member is supported by an oblique strut. The structural members of the base and the track segments are usually made as ladders whose sides constitute a raceway for the carriage's vehicle wheels. The track is typically made of aluminum alloys, but track members made from tubular steel sections are also commercially available.

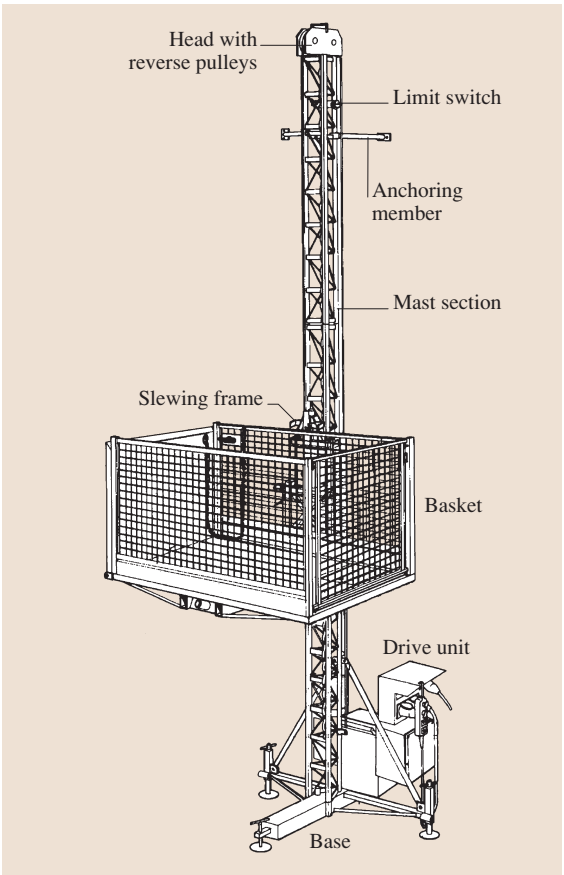


Fig. 14.60 Five-hundred-kilo–capacity lift with cable hoisting drive

Table 14.2 Specifications of cable-driven material and equipment lifts shown in Figs. 14.58 and 14.60

Parameters		
Lifting capacity (kg)	200	500
Lifting speed (m/min)	27	
Maximum lifting height (m)	80	60
Mast section length (m)	2.0	2.0
Mast section weight (kg)	16	30
Distance between anchors (m)	4	4
Gripping device	Yes	Yes
Platform's dimensions		
– Height (m)	max. 1.95	1.0
– Width (m)	0.5–0.75	1.0
– Length (m)	1.15–1.25	1.5
Electric supply (V/Hz)	Single-phase current 230/50	Three-phase current 400/50
Winch motor power (kW)	1.5	3.5
Total mass of lift (kg)	20 m high lift – about 500	20 m high lift – about 775

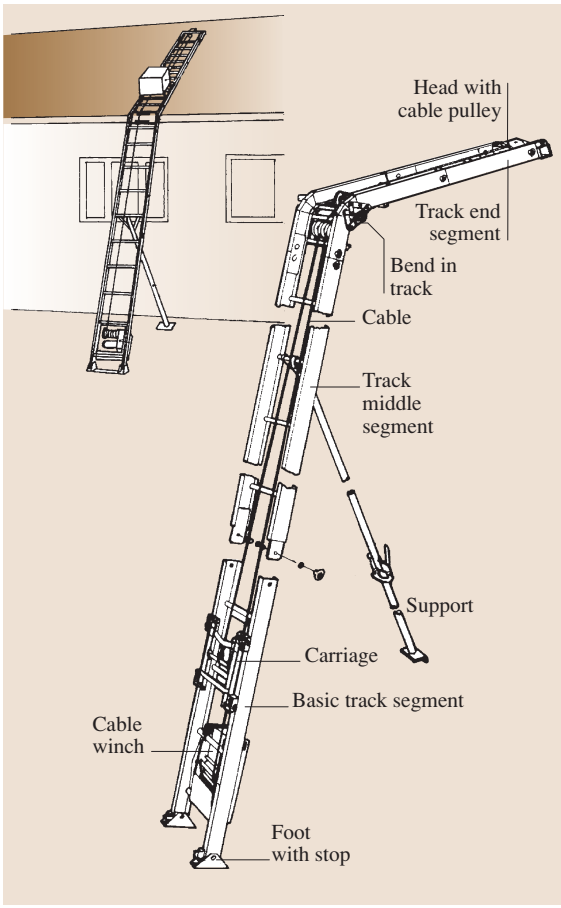


Fig. 14.61 Ladder lift

Within the track structure one can distinguish (Fig. 14.61):

- A base segment
- Middle segments
- A head
- An optional end segment

The base member ends with two hinged feet or a cross-bar with wheels, enabling the ladder lift to be rolled without it being necessary to disassemble it.

The middle segments are equipped with locking connectors inserted into adjoining segments and fixed with bolts and nuts.

The track bend has a link mechanism that allows it to be inclined. It also has cable-guiding pulleys.

Another cable-guiding pulley is situated at the end of the track.

The track's accessories include a strut to support its end passing through a window opening.

In some models, the track can be used as the lift's mast for the vertical transport of materials.

The carriage (Fig. 14.61) is the lift's vehicular unit. It is in the form of a frame with vehicle wheels. A cable with a hook, led from the winch to the carriage, runs along the guide and is directed by intermediate cable pulleys and the pulley in the cable track's head. As a rule, the carriage is equipped with an automatic gripping device actuated when the cable breaks.

Accessories for fastening loads are shown in Fig. 14.62. The accessories are mounted on the carriage. They enable efficient and safe materials handling.

Table 14.3 Specifications of model ladder lifts

Parameters	
Lifting capacity (kg)	150–200
Length of track (m)	13–30
Speed of platform (m/min)	18–40
Guide bend angle adjustment (°)	20–45; in some ladder lifts: 0–70
Number of guide sections	9 to 15
Electric supply (V)	230 single-phase alternating current
Motor power (kW)	0.75–1.5
Control	Electrical – through control buttons
Mass of drive unit (kg)	40–55
Mass of guide section (kg)	≈ 10
Total mass of ladder lift (kg)	160–210

Mast Material and Equipment Lifts with Rack-and-Pinion Hoisting Gear

In material and equipment lifts with a rack-and-pinion hoisting gear the platform used for transporting materials climbs a rack running along the mast.

The lift consists of a platform and a mast with a base. The mast structure is segmental. The individual segments have the form of a cuboidal truss structure with a rack attached to one side. Thanks to its segmental structure and the use of appropriate assembly accessories the lift is a self-assembly device. The mast is extended by adding and joining mast sections. The mast is anchored, usually by means of tubular elements, enabling the adjustment of its position relative to the wall of a building. The drive unit mounted on the platform typically consists of an electric motor with a built-in electromagnetically released spring brake, a flexible clutch, a worm gear, and a rack wheel that mates with the mast's rack. In the case of two-mast lifts, two drive units are employed. Examples of material lifts with a platform raising rack-and-pinion hoisting gear are shown in Figs. 14.63, 14.64 and Table 14.4.

The handled materials are loaded and unloaded at stops with transport stages. The mast's base is enclosed by fencing with a control box attached to it. There is an electrically controlled pivoting gate in the fencing. On the entry side all platform stops have sliding barriers which may be equipped with an electric barrier-opening monitoring system for additional protection, so that persons at the stop cannot fall out. Similarly as ca-

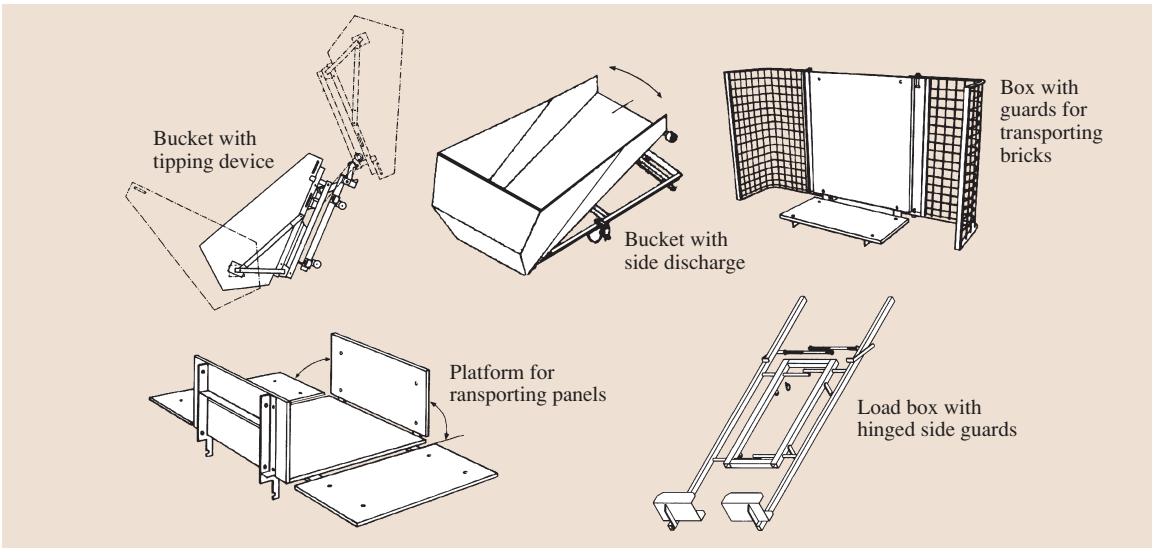


Fig. 14.62 Load fastening accessories

Table 14.4 Specifications of selected material and equipment lifts with rack-and-pinion hoisting gear

Parameters	Single-mast lift	Two-mast lift
Lifting capacity (kg)	600–2000	1500–2000
Maximum lifting height (m)	100	100
Lifting speed (m/min)	About 25	About 25/12
Mast section (m)	1.5	1.5
Height of free-standing mast (m)	9.0	9.0
Distance between anchors (m)	7.5	7.5
Gripping device	Yes	Yes
Platform's dimensions		
– Height (m)	1.1	1.1
– Width (m)	1.5	1.5
– Length (m)	1.5	3.0
Electric voltage (V/Hz)	230–400/50	230–400/50
Hoisting gear motor power (kW)	4.4	2×5.5

ble lifts, lifts with a rack-and-pinion lifting gear may be equipped with a set of wheels to enable them to be easily transported to a new work site. Instead of the platform for handling lump and sacked materials, a special bucket for transporting concrete mix can be installed.

A comparison of lifts with a rack-and-pinion hoisting gear with lifts with a cable hoisting gear shows that the former lifts have several advantageous properties such as: greater lifting height, easier and safer assembly, and simpler operation.

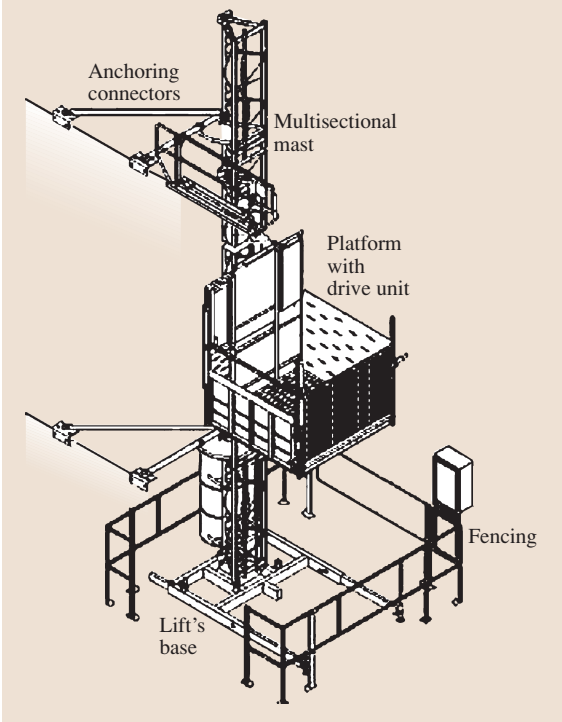


Fig. 14.63 Single-mast material and equipment lift with rack-and-pinion hoisting gear

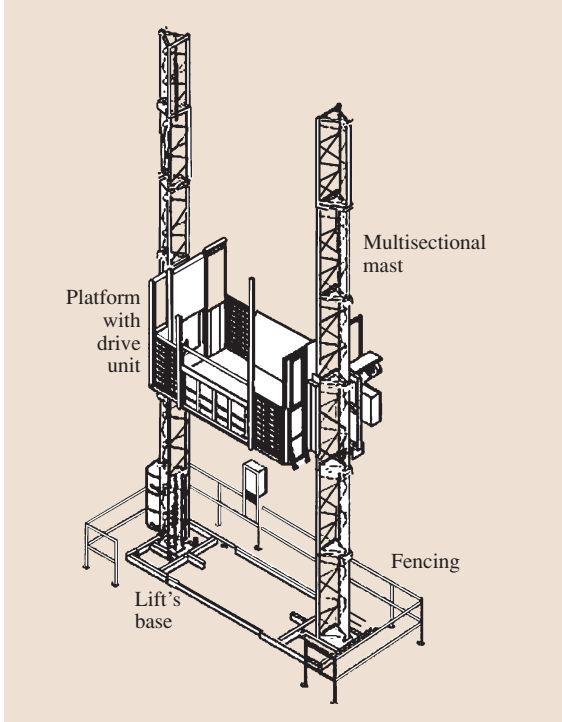


Fig. 14.64 Two-mast material and equipment lift with rack-and-pinion hoisting gear

Table 14.5 Specifications of selected shaft material and equipment lifts

Parameters	
Lifting capacity (kg)	500–1500
Lifting height (m)	15–70
Lifting speed (m/min)	18–33
Platform's dimensions (m)	2×1×1.5
Electric motor	
– Supply voltage and frequency (V/Hz)	230–400/50
– Motor's power (kW)	7–10
Total mass of 15 m- and 70 m-high lift, respectively (kg)	2300–10 000

Shaft Material and Equipment Lifts

Shaft material and equipment lifts (Fig. 14.65) are used for the vertical transport during the construction of medium- and high-rise building structures.

A shaft lift consists of the following main parts (Fig. 14.65 and Table 14.5):

- A shaft
- An upper beam
- Guides
- A bottom cable pulley
- The platform's upper beam
- A platform
- A winch (typical lifting winches with an electric or diesel drive can be used)

The shaft has a spatial truss structure. The load-bearing platform is made of steel sections. The cable is guided by the bottom and top cable pulleys. The end of the cable is fixed to the upper beam (Fig. 14.65).

There is also a shaft lift design in which the shaft is made of only flat frames anchored to the building's wall. Shaft lifts are equipped with similar safety devices as other material lifts.

Material lifts are equipped with the following safety devices:

- A gripping device which stops the platform as it descends whenever it exceeds the maximum allowable rate of descent.
- Protection against disengagement of the drive wheel from the mast's rack. As standard, sliding guides are used. They keep the load platform on the mast even if the roller guides fail.
- An emergency lowering system used in the case of a prolonged power failure. Some lifts are equipped with emergency lowering systems with speed self-stabilization – the speed stabilizes below the speed at which the gripping device is actuated.
- The upper and lower limit switches, automatically stopping the platform at the mast's highest and lowest levels.
- Switches and locks for stop doors or barriers, preventing their accidental opening when the platform is outside the stop zone or in motion.
- Stops to ensure that the platform will be brought to a stop if the limit switches fail.
- An induction sensor that monitors mast presence during mast assembly.
- A sound system signalling the start of a platform ride.
- Protection against electric shock.
- Overload protection of the electric motors.
- Switches actuated when the working platform skews in two-mast lifts.

The operation of the cable-driven material lifts described above typically consists of the control of the movement of the carriage by pushing buttons on the control panel at the lower station. It is also possible to switch to control from the platform during assembly and maintenance of the lift.

14.4.2 Material and Equipment Lifts with Access to Personnel

Material and equipment lifts with access to personnel are intended for the vertical transport of persons and materials during construction/assembly works and repairs of mainly high-rise buildings in housing and industrial construction. Their design is usually similar to that of material and equipment lifts with a rack-and-pinion hoisting gear.

A person and material lift consists of a cabin with a rack-and-pinion drive, moving on a mast secured at the bottom to the lift's base and anchored to the building's wall, and transport stages (stops) between which transport takes place. The mast has a segmental structure and can be extended by adding mast sections. It is anchored by means of a system of tubes, which makes it possible to adjust the mast's position relative to the building's wall.

The lift can have two cabins, each with its own drive system, whereby the transport of persons and materials can be doubled. The lift's cabins move on a common mast independently of each other. Examples of person

Table 14.6 Specifications of selected person and material and equipment lifts

Parameters	Single-cabin lifts	Two-cabin lifts
Lifting capacity (kg)	1200–2000	2×1200–2×2000
Maximum lifting height (m)	150–300	150–300
Maximum number of persons	15–25 depending on lifting capacity	15–25 depending on lifting capacity
Lifting speed (m/min)	≈ 40	≈ 40
Mast section (m)	1.5	1.5
Distance between anchors (m)	12.0–15	12.0–15
Lifting height without anchoring (m)	12.0–15	12.0–15
Davit's lifting capacity (kg)	150	150
Gripping device	Yes	Yes
Cabin's dimensions		
– Height (m)	2.1–2.7	2.1–2.7
– Width (m)	1.3–1.5	1.3–1.5
– Length (m)	≈ 3.0	≈ 3.0
Electric specifications		
– Supply voltage and frequency (V/Hz)	230–400/50	230–400/50
– Motor power (kW)	2×9.0–2×11	2×2×9.0–2×2×11

and material lifts with a rack-and-pinion, platform lifting gear are shown in Figs. 14.66, 14.67 and Table 14.6.

The development of high-rise construction created a need for high-lifting-speed vertical transport equipment. For this purpose fast lifts with a lifting speed of up to 1.8 m/s are employed. These are used for the vertical transport of persons and materials and equipment in industrial construction, for building reinforced concrete chimneys, silos, television towers, and similar structures. One- and two-mast high-speed construction lifts are available.

These lifts incorporate electrohydraulic drive systems whose basic unit is a hydrostatic gear. The hydraulic engine's output shaft is connected by a cou-

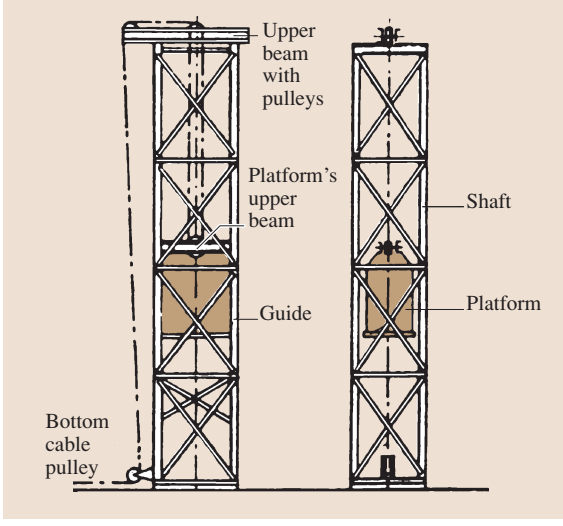


Fig. 14.65 Shaft material and equipment lift with cable hoisting gear

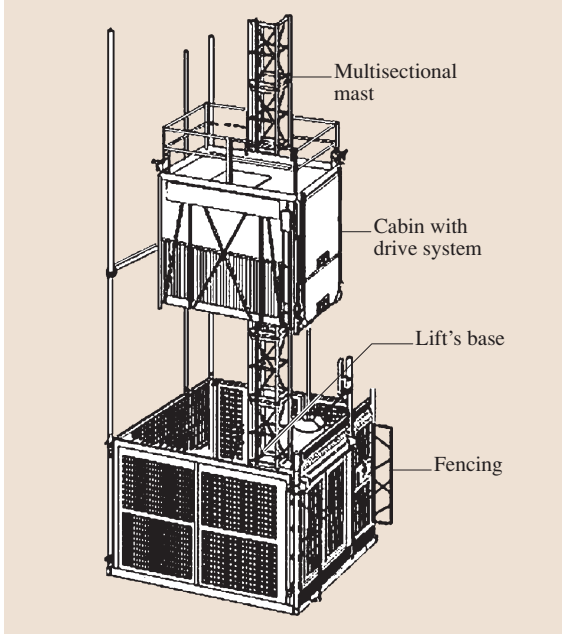


Fig. 14.66 Single-cabin person and material lift with rack-and-pinion hoisting gear

pling to the shaft of a worm gear assembly on the output shafts of which cylindrical gears mating with the mast's rack, and so making the cabin move up or down, are mounted. The weight of the loaded cabin is counterbalanced by a counterweight connected to the cabin by a steel cable passing through pulleys fixed to the top of the mast.

The safety of persons in person and material lifts is ensured by appropriate guards around the traffic way and the stop platform access. Person and material lifts can be controlled from the moving cabin as well as from any stop level. The lift's base (the bottom stop) is fenced in to a height not less than 2.0 m. The access areas are protected by stop doors with a minimum height of 2.0 m, equipped with safety locks.

Entrances to the person and material lifts' cabins are protected by doors. The cabin door is equipped with mechanical bolting devices and safety cutout switches, preventing the door from being opened as the cabin is moving when the cabin's floor is not within about ±0.25 m from a stop landing or when the door is not closed and the bolting device is not in the closed position.

In addition, person and material lifts are fitted with similar safety devices as those used in rack-and-pinion material and equipment lifts, i. e.:

- Gripping devices actuated at an excessive speed of descent
- Protection against disengagement of the drive toothed wheel from the mast's rack
- An emergency lowering system
- Limit switches
- Working platform skewness switches in two-mast lifts
- Protection against electrical failures in the case of no voltage, voltage decay, or voltage drop
- Electrical devices protecting:
  - Closed stop gate position
  - Stop gate bolting device position
  - Closed cabin or platform gate position
  - Bolted emergency hatch or door position

Modern rack-and-pinion person and material and equipment lifts are characterized by:

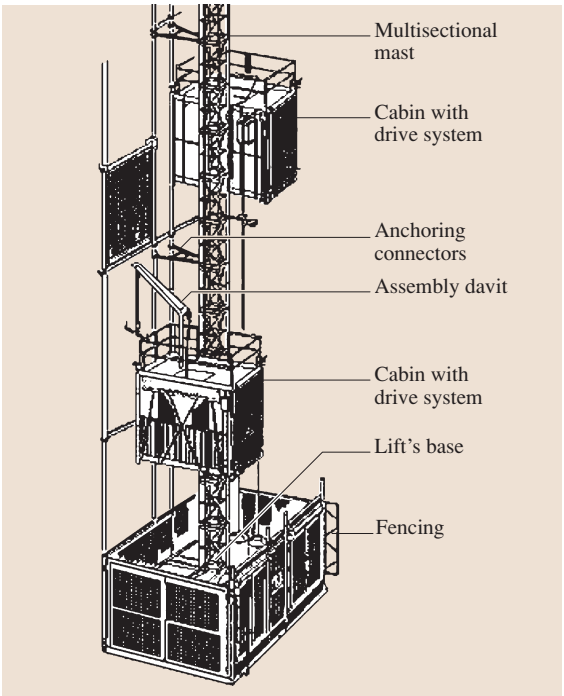


Fig. 14.67 Two-cabin person and material lift with rack-and-pinion hoisting gear

- A great lifting height: up to 300 m
- Easy and quick assembly
- A high lifting capacity: 2000 kg per cabin
- Automatic stopping of the cabin at the terminal stops
- The possibility of programming at which stops the cabin should stop
- Completely safe operation owing to the use of appropriate protective measures
- Easy operation and simple maintenance
- An installation that enables audio communication between the cabin and the bottom stop
- An overload control system

Because of their advantages person and material lifts with a rack-and-pinion hoisting gear have gained a dominant position in the lift market.

14.5 Access Machinery and Equipment

14.5.1 Static Scaffolds

A scaffold is a temporary (usually bar) structure erected to provide safe access during construction, repair, maintenance, and demolition of all kinds building structures.

Scaffolds can be classified according to the different criteria specified in Table 14.7, and the main criteria for classifying scaffolds and the related terminology are described below.

One of the major criteria is the division of scaffolds with regard to their design and assembly method. Particularly important is the division into tube–coupler scaffolds and system scaffolds.

Tube–coupler scaffolds (Fig. 14.68) are constructed from steel tubes and couplers, and the stagings are made from boards or balks. In this type of scaffolds, the dimensions of the structural grid are not rigidly imposed by the dimensions of the components, e.g., by the length of the tubes. Workers assembling a tube–coupler scaffold according to a blueprint ascertain the positions of all the elements which determine the dimensions of the structural grid and the verticality of the uprights. The basic components of the tube–coupler scaffold are shown in Fig. 14.68.

In tube–coupler scaffolds, such elements as standards, transons, and bracings are joined eccentrically by right-angle or swivel couplers, as illustrated in Fig. 14.69.

A characteristic feature of *scaffolds made of prefabricated elements* (system scaffolds) is that their dimensions (or some of their dimensions) are determined by the dimensions of their components. All frame and modular scaffold systems belong to this class. A general view of system scaffold constructions is shown in Fig. 14.70. Modular scaffolds and frame scaffolds are shown on the left and right, respectively.

In the *frame scaffold*, the vertical structure is made up of prefabricated flat frames. The frame consists of two uprights permanently connected by transverse el-

Table 14.7 Classification of scaffolds

No.	Classification criterion	Name of scaffold
1	Intended use	Working scaffold Protective scaffolds Load-bearing scaffolds
2	Design and assembly method	Tube–coupler (bricklayer's) scaffolds Ladder scaffolds Scaffold made of prefabricated elements (system scaffold): Modular, Frame
3	External load-bearing mode	Standing scaffolds Suspended scaffolds Trestle scaffolds Outrigger scaffolds Cantilever scaffolds
4	Protection of scaffold against overturning	Anchored facade scaffolds Scaffolds secured to base by guy-ropes Free-standing scaffolds
5	Transferability	Immobile (stationary) scaffolds Mobile (portable) scaffolds
6	Operating mode	Scaffolds used for short periods, e.g., during working shift Permanent scaffolds used for prolonged periods without dismantling
7	Material from which scaffold load-bearing elements are made	Wooden scaffolds Aluminum scaffolds Steel scaffolds
8	Technical–organizational and formal-legal aspects	Scaffolds in typical version Individually designed scaffolds

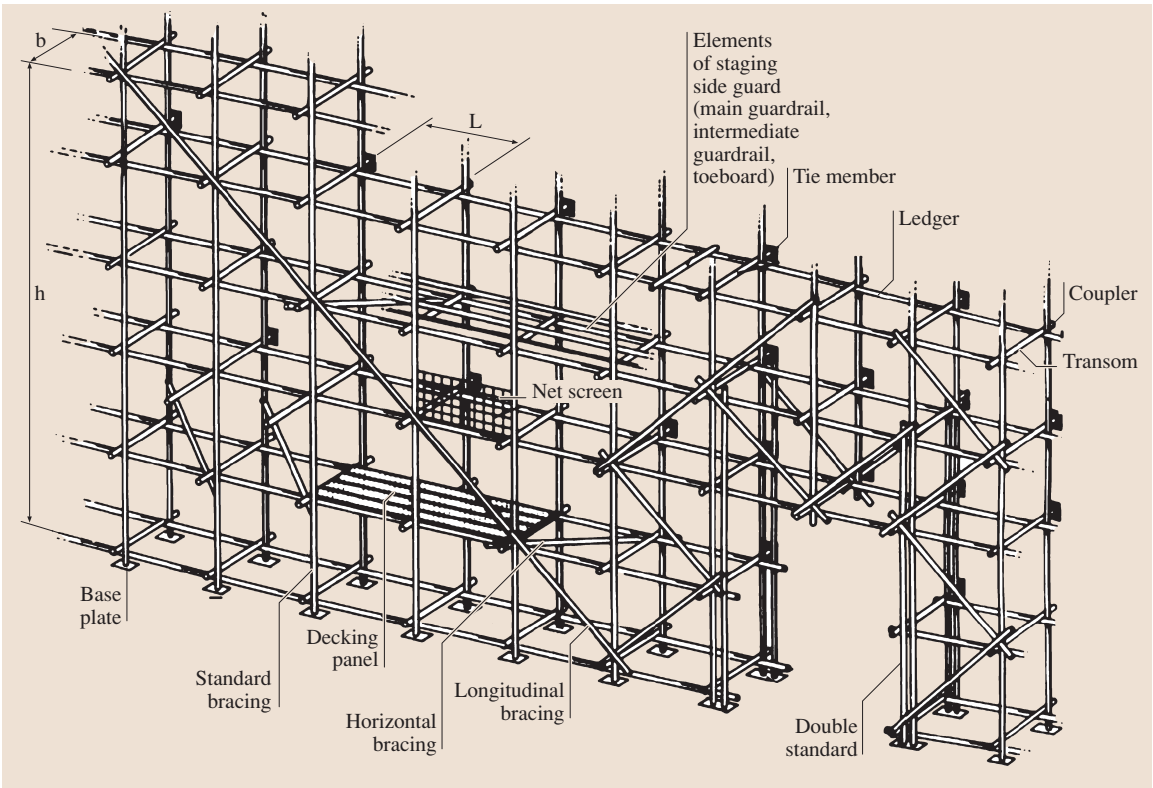


Fig. 14.68 Anchored facade tube–coupler scaffold (*b* scaffold width; *L* scaffold length; *h* scaffold height)

ements at the frame's top and bottom. The top cross member is used for fixing decking panels. The bottom member may prevent the disassembly of the decking panel while the scaffold is in service. In the vertical plane the scaffold is braced by diagonal bracings. In the horizontal plane it is braced by system decking panels.

In a modular scaffold, transoms, ledgers, and bracings are joined with standards at fixed nodal points spaced at regular intervals, usually every 0.5 m.

Modular and frame scaffolds made by one manufacturer are in most cases compatible and can be combined.

In frame system scaffolds, uprights are connected with crossbars by inseparable welded joints.

In *modular system scaffolds*, standards are joined with transoms at fixed nodal points.

Transoms and bracings can be attached to standards by means of coupling elements permanently fixed to the latter. The coupling elements (coupling heads) usually have the form of a disk or a flange and they are regularly spaced, usually every 0.5 m. The bracings and the transoms are tipped with coupling heads. Transoms and bracings are connected with standards, typically by cotter joints. As opposed to tube–coupler structures, in modular scaffolds the axes of the ledgers, the transoms, and the standards connected together intersect at one point. A modular scaffold joint is shown in Fig. 14.71.

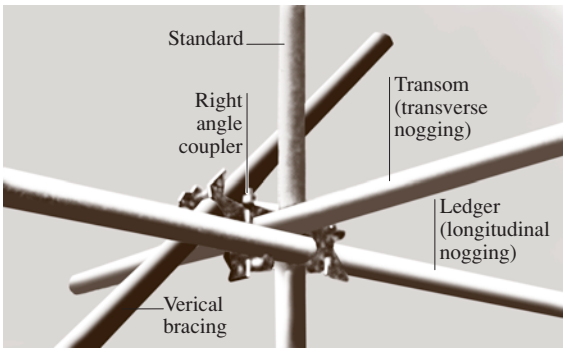


Fig. 14.69 Tube–coupler scaffold joint

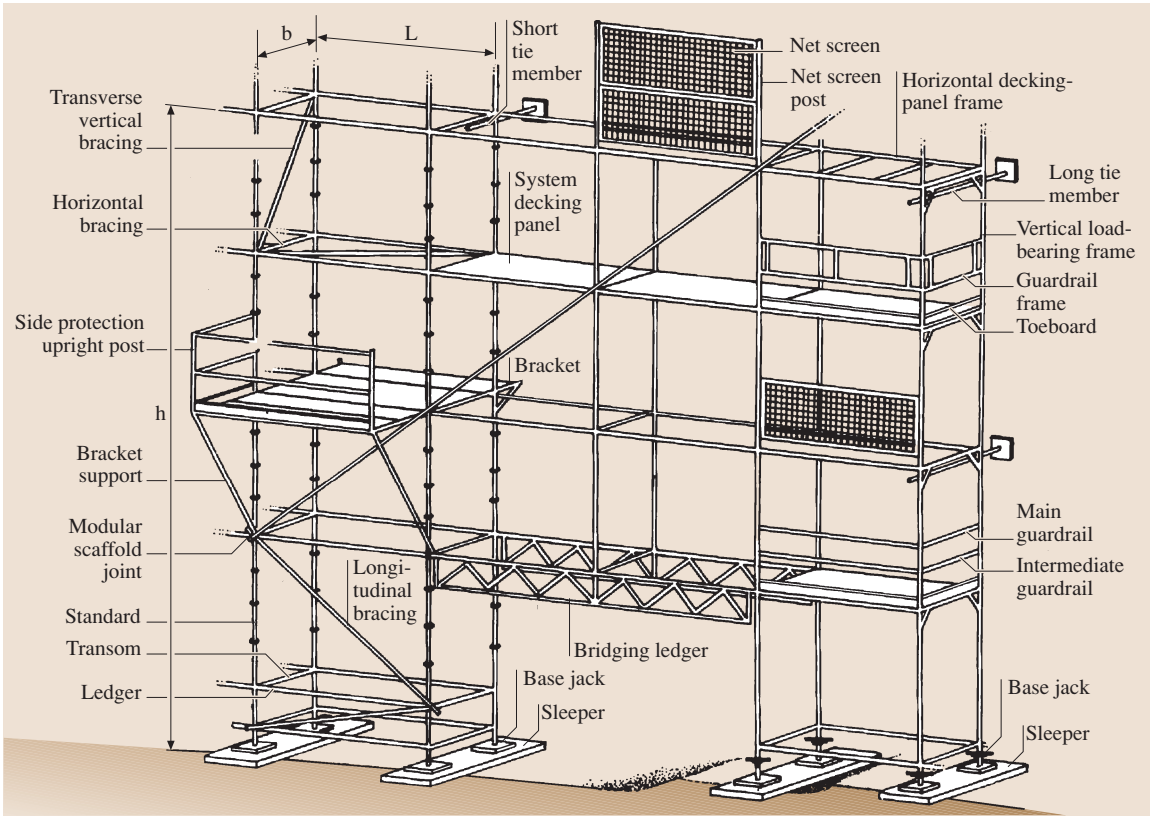


Fig. 14.70 Anchored facade system scaffolds frame scaffold (*right*) and modular scaffold (*left*) (*b* scaffold width; *L* scaffold length; *h* scaffold height)

System scaffolds have several advantages over tube–coupler scaffolds:

- Much easier, quicker, and safer assembly
- Higher load capacity at similar or identical scaffold geometrical or material specifications
- Lower and easier to assess variability of the scaffold structure's random parameters such as geometric imperfections, the load capacities of the individual elements, and the characteristics of the joints
- Possible greater standardization of typical designs

For these reasons tube–coupler scaffolds are being supplanted by system scaffolds.

Another major criterion according to which scaffolds are divided is their intended use. This is connected with the character and magnitude of the loads acting on the assembled structure. Two basic uses can be distinguished:

- Working scaffolds – structures capable to carry worker, material, and equipment loads.
- Load-bearing scaffolds – support structures which, during the construction of a building, can be loaded with the weight of its individual elements or units; one should note that the reliability of scaffolds of this type is a necessary condition for the proper course of a construction process (e.g., concreting). The construction and use of load-bearing scaffolds is subject to separate regulations.

Another major scaffold classification criterion is connected with technical–organizational and formal–legal aspects and with the standardization of scaffold designs and applications.

In legal instruments scaffolds are divided into:

- Scaffolds in a typical version
- Individually designed scaffolds

Table 14.8 Specifications of typical portable single-mast climbing platforms

Parameters				
Platform's length/lifting capacity (m/kg)	4.1/1300 7.1/800 10.1/500	4.2/1300 7.4/1000 10.5/700	4.2/2000 7.4/1700 10.5/1400 12.5/1200	4.2/2700 7.3/2300 10.5/1900 13.7/1500 16.9/1000
Maximum platform elevation without anchoring				
– Protractible beams protracted on both sides (m)	6	20	15	18–20
– Protractible beams protracted on one side of mast (m)	6	15	15	13–15
Maximum platform elevation with one mast anchoring point located at top (m)	11.5	25	25	25
Maximum platform elevation with anchoring along entire length of mast (m)	100	200	200	200
Spacing between anchors (m)	6	12.5	12.5	12.5
Max. length of platform's protractible part (m)	1.0	0.3	1.4	2.5
Max. loading of struts (kN)	15	50	60	65
Lifting speed (m/min)	6	6	6	6
Transport mass (kg)	1800	3500	4000	4000
Mast section: length/weight (mm/kg)	1508/48	1256/82	1256/82	1256/82
Electric specifications of platform lifting gear	400 V/50 Hz 3 kW, 16 A	400 V/50 Hz 3 kW, 16 A	400 V/50 Hz 3 kW, 16 A	2 × 400 V/50 Hz 3 kW, 16 A

A scaffold in a typical version means an assembly version of the scaffold which covers the most frequent applications of the scaffold structure. It is assumed that the manufacturer has provided a proof of the scaffold's static strength and neither its user nor the company assembling it has to provide such a proof in order to certify the scaffold fit for use on the construction site. Also in the case when the assembly version has been realized in accordance with a generally recognized assembly standard the proof does not have to be provided. The generally recognized assembly standard may be defined in assembly norms or instructions issued by the manufacturer of the given type of scaffold.

Scaffolds in a typical version are anchored, facade, working scaffolds with a height of up to 24 m and access working towers erected to a height of 8 m outdoors and to a height of 12 m indoors.

Typically, a useful load of 2 kN/m^2, sometimes 0.75, 1.5, 3, 4.5, and 6 kN/m^2, is assumed for stagings.

Examples of scaffolds in their typical versions are shown in Fig. 14.72.

Each scaffold that is not in a typical version should be individually designed and its statics tested. The range of the structural analysis depends on the complexity of a given scaffold structure. Examples of atypical scaffolds are shown in Figs. 14.73–14.77.

Competent construction design companies, usually connected with equipment manufacturers, should be entrusted with the design of atypical scaffolds. When selecting scaffolds the user should take into account the following:

- The construction site's location (a wind load zone, power lines, traffic routes, etc.)
- The kind of terrain and the lay of the land on which the scaffold is to be founded

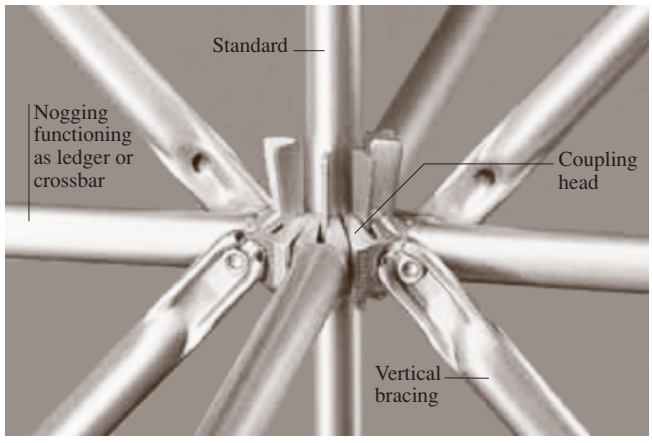


Fig. 14.71 Modular scaffold joint

- The bearing capacity of the base on which the scaffold is to be founded
- The building's height and the shape of the elevation
- The kind and type of scaffold
- The intended use of the scaffold
- The magnitude of the loads originating from people and equipment
- Easy anchoring of the scaffold
- The number and layout of traffic routes
- The transport of materials onto the scaffold
- Canopies

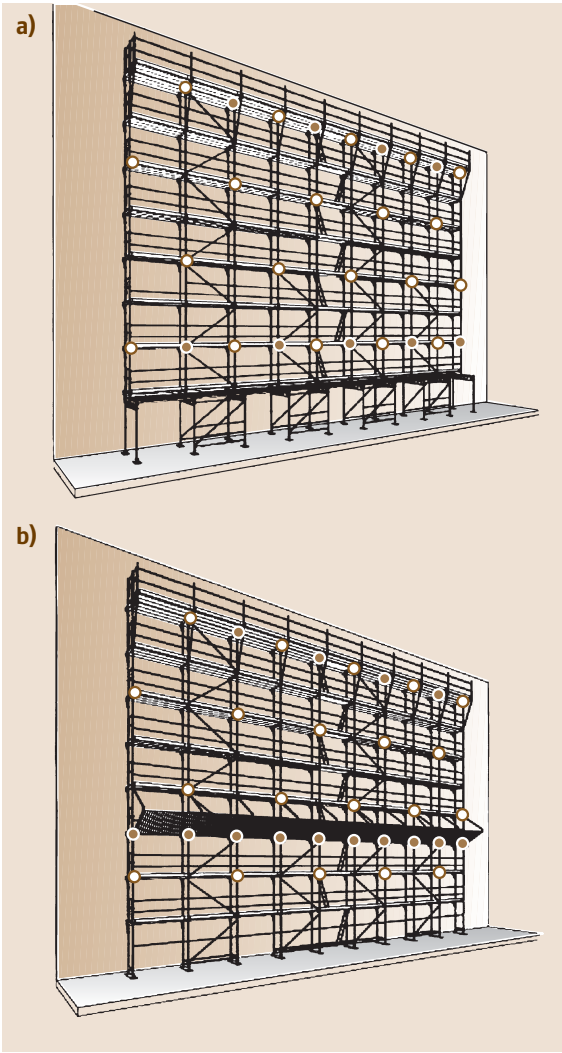


Fig. 14.72a,b Example of anchored, facade, frame scaffold in typical version: **(a)** version with first story constructed from intermediate frames; **(b)** version with canopy

- The assembly of a typical or an atypical scaffold
- Renting or purchasing costs

14.5.2 Elevating Work Platforms

General

Elevating work platforms form a wide class of cranes for elevating persons and equipment for repair, maintenance, and assembly purposes. Depending on their design and application, elevating work platforms can be divided into the following groups:

- Portable mast-climbing platforms
- Mobile (mounted on special chassis) elevating work platforms
- Hanging scaffolds

Portable Mast-Climbing Platforms

Portable mast-climbing platforms are used for masonry plastering, assembly, insulation, and facade works in housing and industrial construction. Portable mast-climbing platforms (Figs. 14.78 and 14.79 and Tables 14.9 and 14.10) are made up of the following units:

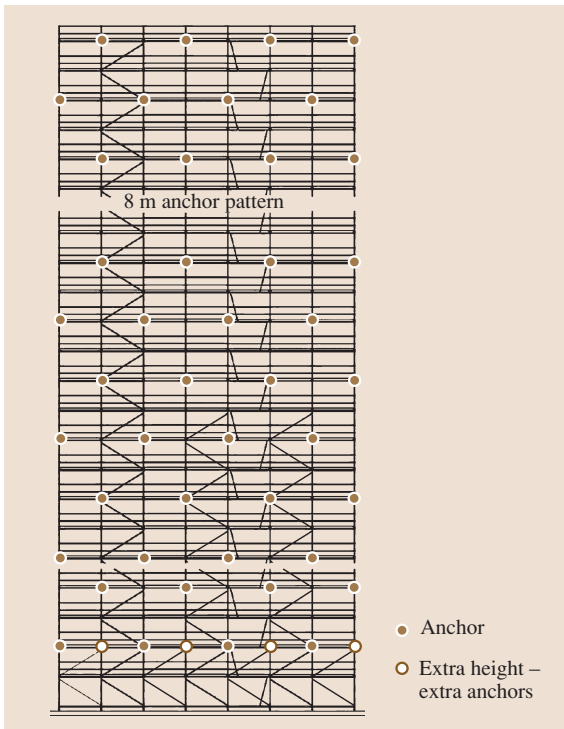


Fig. 14.73 Anchored, facade, frame scaffold in atypical version. The atypicality consists of the considerable height of the scaffold

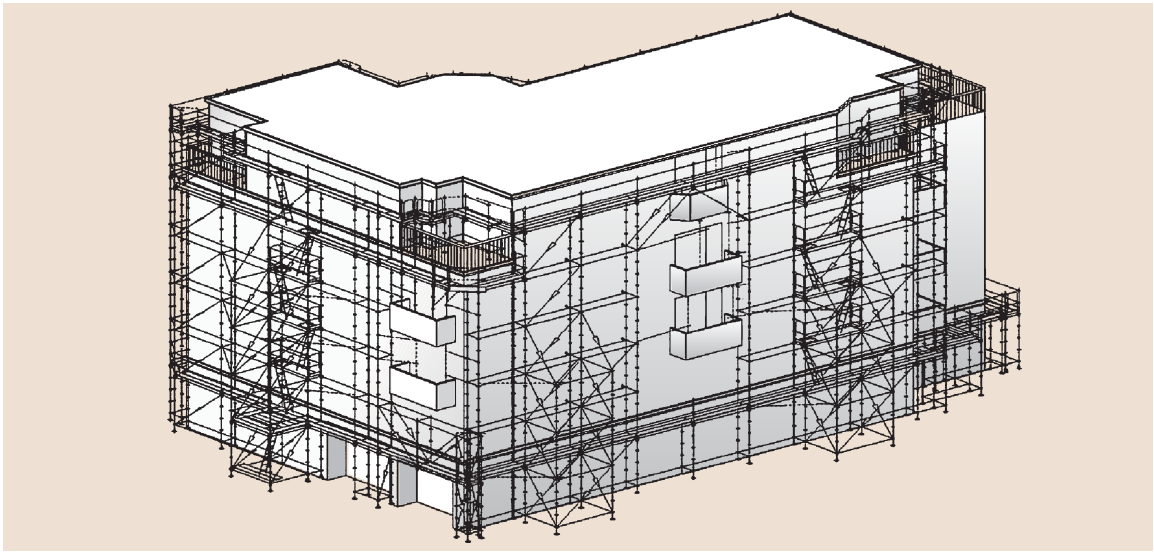


Fig. 14.74 Facade, frame scaffold in atypical version. The atypicality consists of the lack of anchoring

- A base
- A mast or multisectional masts
- A work platform lifting unit
- A work platform

The work platform moves on a mast fixed to a base and anchored to a wall of a building structure. The mast is

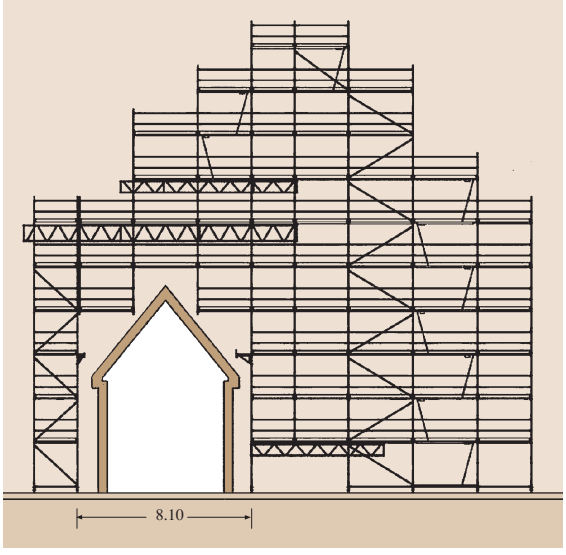


Fig. 14.75 Anchored, facade, frame scaffold in atypical version. The atypicality consists of the part of the scaffold suspending on girders

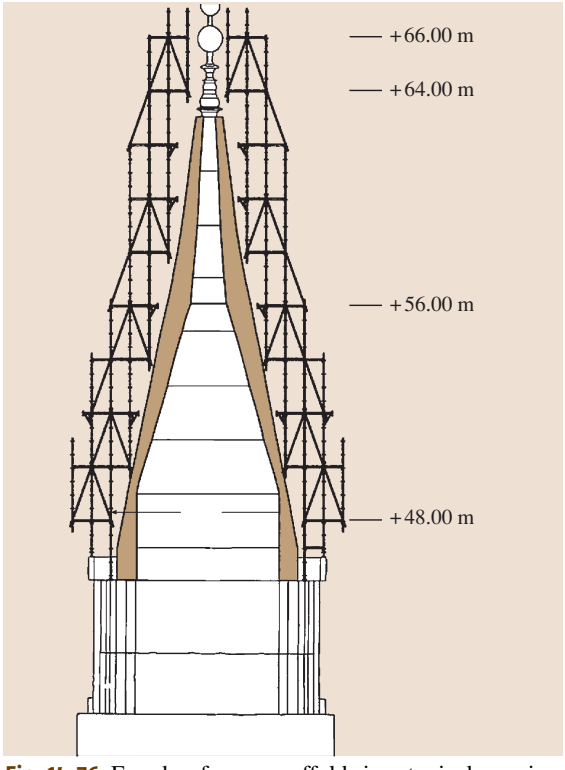


Fig. 14.76 Facade, frame scaffold in atypical version. The atypicality consists of the shape of scaffold support scheme, which differs greatly from the typical version

Table 14.9 Specifications of typical portable two-mast climbing platforms

Parameters				
Platform's length/lifting capacity (m/kg)	4.1/1300 7.1/800 10.1/500	4.2/1300 7.4/1000 10.5/700	4.2/2000 7.4/1700 10.5/1400 12.5/1200	4.2/2700 7.3/2300 10.5/1900 13.7/1500 16.9/1000
Maximum platform elevation without anchoring				
– Protractible beams protracted on both sides (m)	6	15	10	20
– Protractible beams protracted on one side of mast (m)	6	15	15	12.5–17.5
Maximum platform elevation with one mast anchoring point located at top (m)	11.5	25	25	25
Maximum platform elevation with anchoring along entire length of mast (m)	100	200	200	200
Spacing between anchors (m)	6	12.5	12.5	12.5
Max. length of platform's protractible part (m)	1.0	0.3	1.4	2.5
Max. loading of struts (kN)	15	50	60	65
Lifting speed (m/min)	6	6	6	6
Transport mass (kg)	3700	2×3500	2×4000	2×4000
Mast section: length/weight (mm/kg)	1508/48	1256/82	1256/82	1256/82
Electric specifications of platform lifting gear	400 V/50 Hz 3 kW, 16 A	400 V/50 Hz 3 kW, 16 A	400 V/50 Hz 3 kW, 16 A	2×400 V/50 Hz 3 kW, 16 A

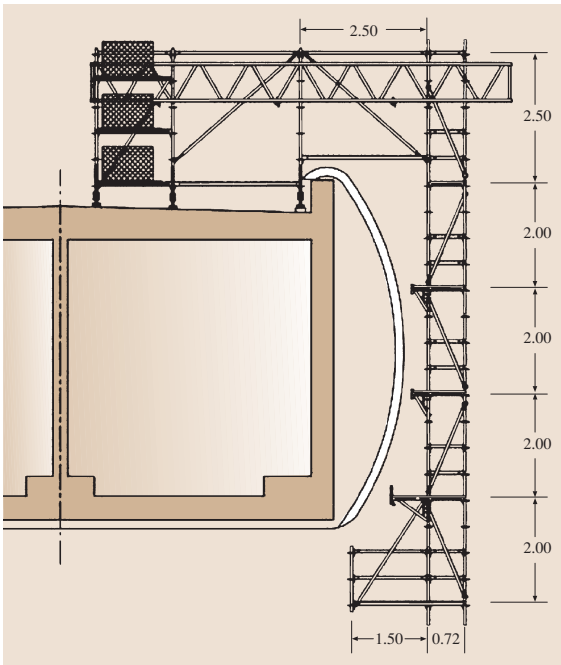


Fig. 14.77 Atypical suspended scaffold. The atypicality consists of the suspending main part of scaffold

made up of sections added up to the required height. In the case of a free-standing mast, the work platform can be usually lifted to a height of 20 m. The mast's base can have the form of a carriage or be stationary. If a carriage is used as the base, the free-standing mast-climbing platform can be moved without it being it necessary to disassemble the mast completely. The carriage can be towed by a tractor or be self-propelled and so able to move along the wall of the building structure. A small-sized stationary base is used when the space for the mast-climbing platform is restricted and no carriage can be employed, e.g., in a street with a narrow pavement. The work platform is elevated by a rack-and-pinion gear. Thanks to its sectional design the mast-climbing platform can be configured to fit the building structure's shape.

In addition, protractible struts, with length adjustable from 0 to 2500 mm, can be attached to the mast-climbing platform's sections. Planks or wooden boards are placed on the beams, thereby creating additional working surface. The combination of the mast-climbing platform's sectional structure and the system of protractible struts makes it possible to obtain work access on walls of any shape (straight, curved, and with slants and bevels) and architectonic form (balcony,

loggia, niche, bay). The whole mast-climbing platform is fenced in with railings to protect persons working on it from falling out. It is also possible to combine two single-mast climbing platforms to form one platform (up to 40 m long) climbing two masts (Fig. 14.79 and Table 14.10). In many cases, mast-climbing platforms may replace stationary construction-assembly scaffolds.

Similarly to material hoists and person and material hoists with a rack-and-pinion drive, mast-climbing platforms are equipped with the following safety devices:

- An emergency lowering system
- A braking device
- A safety device preventing the driving gear wheel from disengaging from the mast's gear rack
- Electric-shock protection
- Overload protection for the electric motors
- Work-platform-slanting cutouts (in two-mast climbing platforms)
- Work platform terminal position cutouts
- Sensors signalling a platform loading which may result in overturning of the mast-climbing platform or its damage

Mobile Elevating Work Platforms

Mobile elevating work platforms have a similar range of applications (elevating persons and equipment) as the portable mast-climbing platforms described above, except that their use in one work place is short.

Mobile elevating work platforms form a class of devices varied in their design. The basic types of mobile elevating work platforms are listed in Table 14.10.

Mobile elevating platforms are usually equipped with protractible struts. Truck-mounted platforms are the most popular mobile elevating work platforms used in construction.

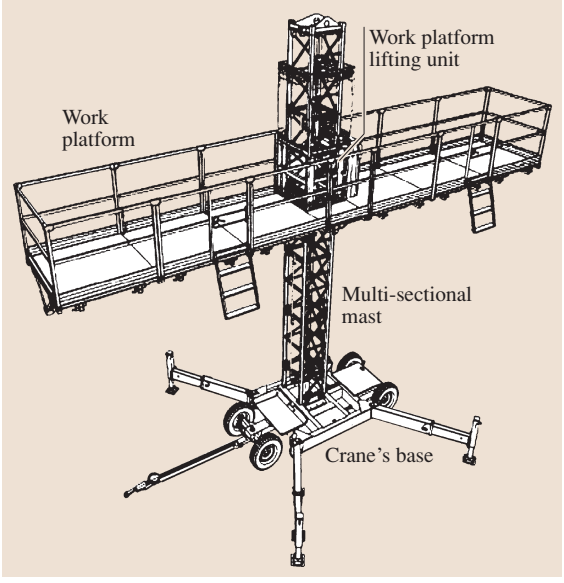


Fig. 14.78 Portable single-mast climbing platform with rack-and-pinion lifting gear

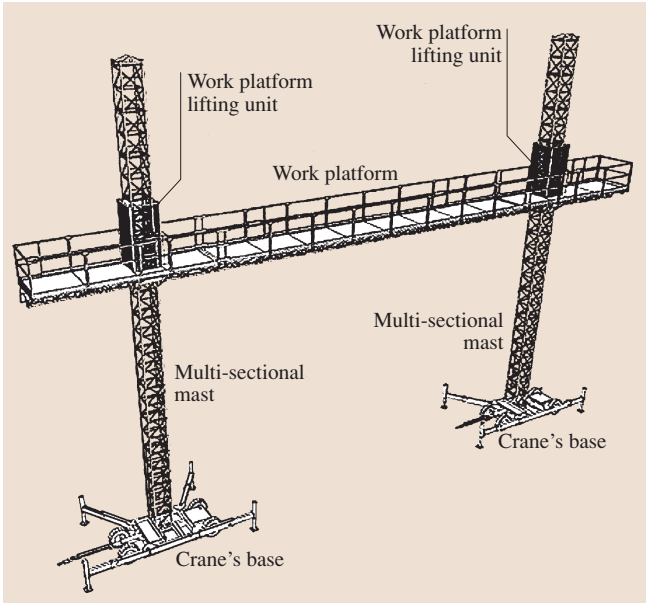


Fig. 14.79 Portable two-mast climbing platform with rack-and-pinion lifting gear

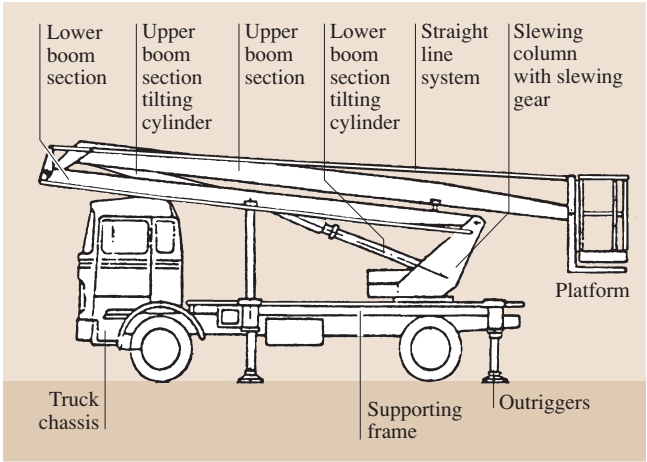
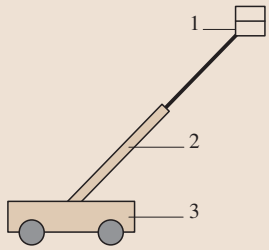
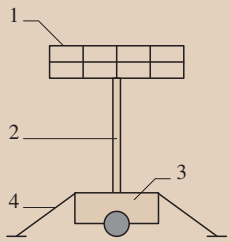
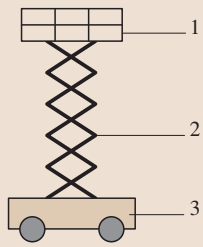
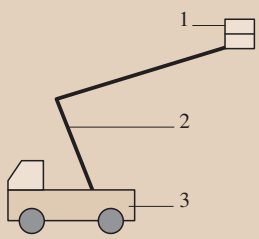


Fig. 14.80 Truck-mounted elevating platform

Table 14.10 Types of mobile elevating work platforms

Description of mobile elevating work platforms	Design schematics
A mobile elevating work platform featuring a telescopic boom. The boom can be raised, lowered, and slewed relative to the vertical axis.	
A platform mounted on a telescopic column or a hydraulic servomotor. The working motions are: the raising and lowering of the platform in the vertical plane.	
A mobile elevating platform with a scissor extending structure.	
A truck-mounted elevating platform whose extending structure is usually in the form of a two-stage boom. The elevating platform performs the following motions: • Raising and lowering • Slewing relative the vertical axis perpendicular to the base	
Key: 1 platform; 2 extending structure; 3 chassis; 4 struts	

Truck-Mounted Elevating Platform

The truck-mounted elevating platform consists of the following parts shown in Fig. 14.80.

The supporting frame is a body to which protractible struts, a hydraulic feeder, and a slewing gear are fixed. The supporting frame is secured to a truck chassis.

Table 14.11 Specifications of selected truck-mounted elevated platforms with elevation up to 15, 21, and 40 m

Specifications	Platforms with elevation of 15 m	Platforms with elevation of 21 m	Platforms with elevation of 40 m
Max. working elevation (m)	10.0–15.0	16.0–21.0	25.0–40.0
Max. horizontal radius (m)	4.2–7.4	5.5–11.0	11.6–21.9
Lifting capacity (kg)	120–200	200–300	265–365
Weight (without chassis) (kg)	950–2300	1620–5900	5200–13 400
Length in transport mode (m)	5.3–7.45	6.8–10.2	7.8–11.4
Width in transport mode (m)	1.9–2.1	2.1–2.5	2.5
Height in transport mode (m)	2.0–3.6	2.6–3.35	3.5–3.8
Slewing angle (°)	330 or 360	360	360

The slewing column is attached to the supporting frame through a crown-bearing. The column is a slewing welded-construction frame with a transmission gear and a brake mounted on it. The lower boom stage and the cylinder are attached to the column by articulated joints.

The boom consists of two stages connected by an articulated joint. Each stage is a welded box structure.

The work platform consists of a floor made from thin-walled steel sections, railings, and curbs.

The straight-line system is a tension-member structure which keeps the work platform in a horizontal position regardless of the angles at which the boom's stages are positioned.

The work platform positioning system consists of a pump, filters, distributors, valves, steel pipes, hoses, and hydraulic cylinders. The hydraulic pump is powered from the truck's gearbox via a lay shaft. Oil is supplied to the working cylinders through a rotary joint. The struts are controlled through distributors mounted on the supporting frame. The boom's cylinders and slewing can be controlled through a distributor mounted near the slewing column or on the work platform. The system is protected against the eventuality of simultaneous steering of the struts and the boom's cylinders. The boom's cylinders and those of the struts are equipped with valves to prevent a pressure failure and the resulting uncontrolled shift of the piston rod.

The maximum angle of elevation of the boom's upper stage is limited by a limit switch in the form of a hydraulic valve controlled by a cam on the boom. The hydraulic system's circuits are protected against excessive pressure rise by overflow valves.

In elevating platforms capable of large elevations (20–40 m), proportional control, ensuring that the working motions controlled by the operating lever are fluid and quick, is used instead of the typical hydraulic control.

Truck-mounted elevating platforms have the following safety measures:

- Limit switches preventing the boom's upper section from being excessively raised and the boom's lower stage from being lowered while the boom's upper stage is maximally raised
- A hydraulic system lock preventing the working circuit and the struts from being simultaneously fed
- Overload protection in the form of overflow valves

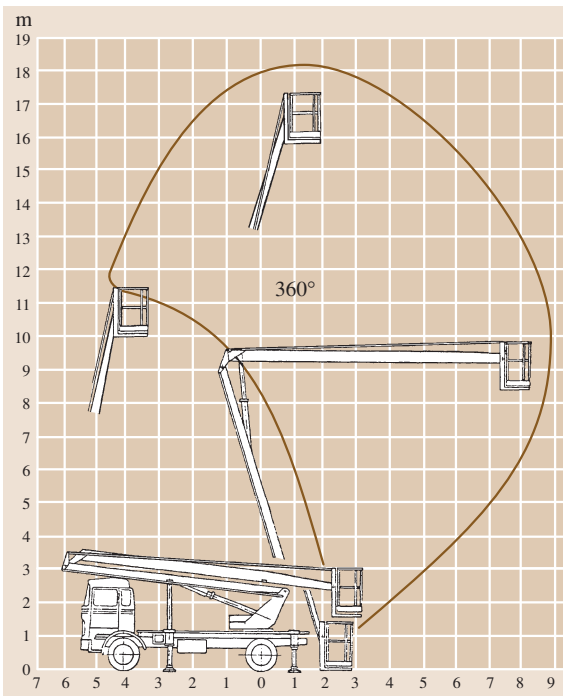


Fig. 14.81 Work area of truck-mounted elevating platform with elevation of 18 m and radius of 8.7 m

- Emergency lowering of the cradle while the pump drive is switched off

The truck-mounted elevating platform's basic operating parameters are:

- Lifting capacity
- Maximum elevation
- Maximum radius
- Work area, specifying the allowable position of the elevating platform in the vertical plane
- Angle of rotation of the body

The work area with specified positions of the work platform in the vertical plane is shown in Fig. 14.81.

Mobile elevating platforms can be mounted on mass-produced truck chassis. The type and size of chassis depends on the specifications of the elevating platform. The characteristics of selected chassis for the particular groups are detailed in Table 14.11.

14.5.3 Hanging Scaffolds

Hanging scaffolds are intended for use in both housing and industrial construction. They are used for finishing works such as painting work, plaster work, insulation work, and window glazing. In many cases, hanging scaffolds may replace stationary construction-assembly scaffolds.

Hanging scaffolds can be divided into the following kinds:

- Stationary, hand-driven, single-person (cradle) scaffolds
- Stationary, hand- or electrically driven scaffolds
- Mobile, hand- or electrically driven scaffolds
- Stationary, hand- or electrically driven sectional scaffolds

Stationary, single-person hanging scaffolds are cradles with a single-person workstation. The cradle is suspended by a hoisting cable from a boom placed on the roof of a building. The vertical motion of the scaffold is effected by means of a two-crank hand winch operated by the person in the cradle. There are containers for materials and tools on both sides of the cradle.

A single-person hanging scaffold is shown in Fig. 14.82.

Specifications of single-person hanging scaffolds are listed in Table 14.12.

Table 14.12 Typical specifications of single-person hanging scaffold

Parameters	Value
Hoisting capacity (kg)	100
Elevation (m)	35
Hoisting speed (m/min)	1.35–2.2
Type of drive	Hand
Crank force (N)	150
Diameter of steel cable (mm)	8
Mass of movable part (kg)	148
Mass of counterweight (kg)	270

Stationary Hanging Scaffolds

These are scaffolds in which the work station is a platform with winches, suspended on cables. Hand-driven and electrically driven hanging scaffolds are used.

A stationary hanging scaffold with a hand drive is shown in Fig. 14.83.

The hanging scaffold shown in Fig. 14.83 consists of three main units: a platform, two winches with a hoisting cable, and a boom. The platform is a steel frame (lined with boards) with a take-down guardrail. The platform is equipped with roller fender beams guiding it on the building's wall during hoisting and lowering. The hanging scaffold is hoisted and lowered

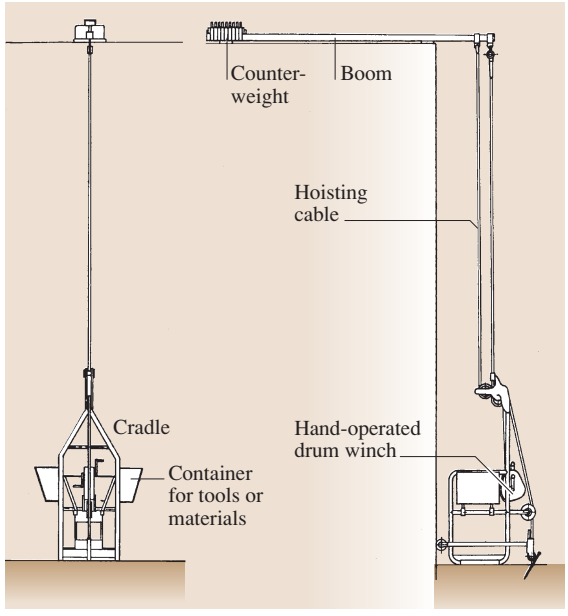


Fig. 14.82 Single-person hanging scaffold

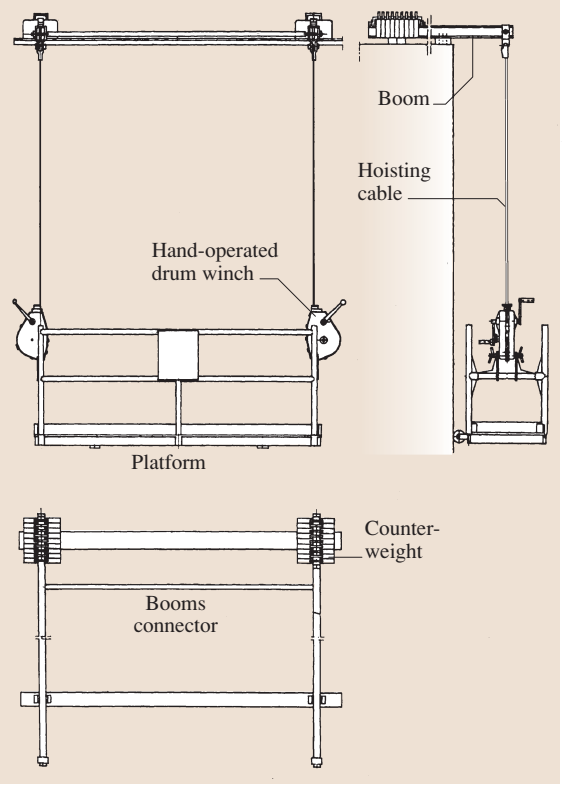


Fig. 14.83 Hanging scaffold with hand drive

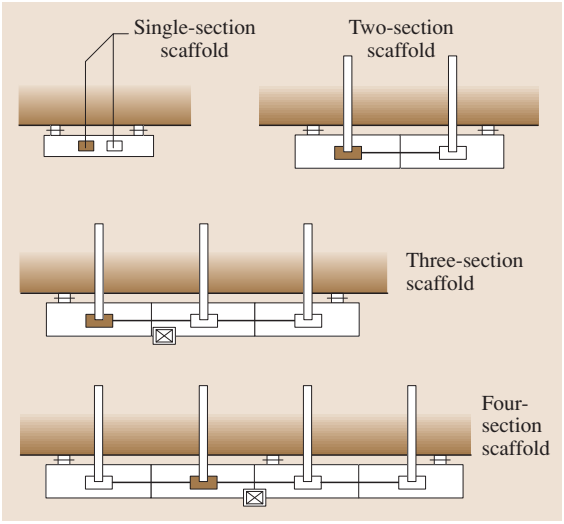


Fig. 14.84 Schematics of sectional hanging scaffolds

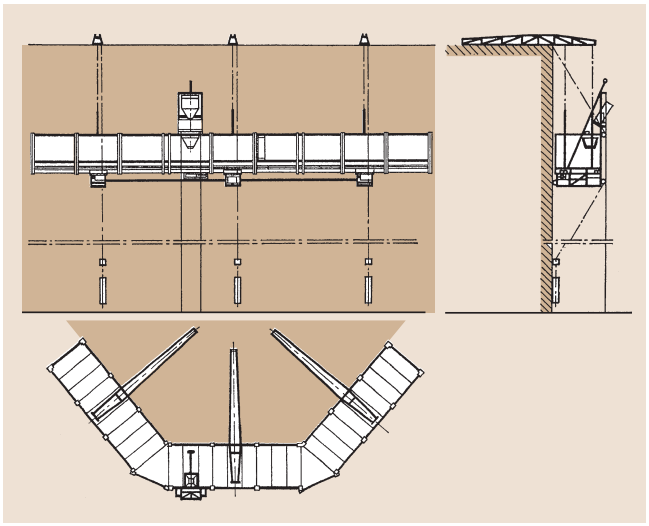


Fig. 14.85 Arched sectional scaffold. The vertical motion of the scaffolds is effected by electric motor cardan shafts and winches

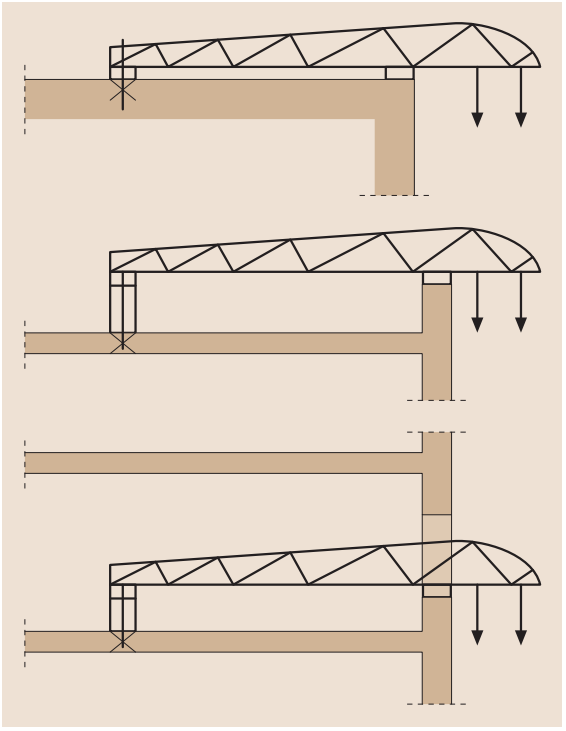


Fig. 14.86 Ways of anchoring sectional scaffolds' booms

Table 14.13 Typical specifications of hand- and electrically driven hanging scaffolds

Parameters	Hand-driven scaffolds	Electrically driven scaffolds
Hoisting capacity (kg)	300	500
Elevation (m)	up to 35	up to 80
Hoisting speed (m/min)	1.35–2.2	4–8
Crank force (N)	150	
Hoisting parameters of electric motor	–	2.2 kW (230/400 V)
Scaffold's dimensions		
Length (m)	3.0	3.1
Width (m)	0.8	1.0
Diameter of steel hoisting cable (mm)	8	11
Mass of movable part (kg)	125	620
Mass of counterweight (kg)	270	–

Table 14.14 Typical specifications of sectional hanging scaffolds with electric drive

Parameters	Scaffolds			
	1-section	2-section	3-section	4-section
Hoisting capacity (kg)	1200	1200	1200	1200
Elevation (m)	80	80	80	80
Hoisting speed (m/min)	4	4	4	4
Scaffold dimensions				
Length (m)	4.5	9.0	13.5	18.0
Width (m)	1.5	1.5	1.5	1.5
Installed power (kW)	5.5	5.5	5.5	5.5
Total weight of scaffold (kg)	900	1770	2500	3610

by means of two hand winches suspended on suspension rods and hoisting cables. The following winch types are used for hand driving hanging scaffolds:

- Typical drum winches with the end of the cable fixed to the drum
- Frictional winches with loading of the lower end of the cable by means of a weight or spring load
- Lever pull hoists

In drum winches a pawl on the winch's ratchet and a frictional mechanism prevent the scaffold from unintended descent. The scaffold is suspended from two booms joined together by pipe connectors. The scaffold is for two persons, each operating one winch during the hoisting and lowering of the platform.

A stationary hanging scaffold with an electric drive is suspended in a similar way to the hand-driven scaffold. The vertical motion of the scaffold is effected by two drum winches with gears. The turns of the winches are synchronized through a cardan shaft connection between the driving electric motor and the winches. Scaffolds of this type are equipped with a hand-operated lowering gear used in the event of failure of the electric drive.

The specifications of hand- and electrically driven hanging scaffolds are shown in Table 14.13.

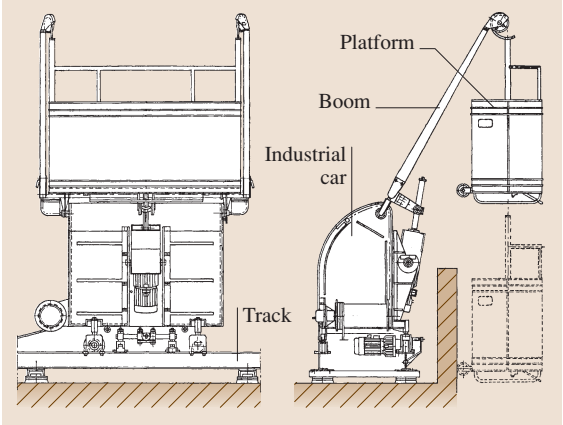


Fig. 14.87 Mobile hanging scaffold with electric drive

Stationary Sectional Hanging Scaffolds

This group includes: straight, angular, and arched hanging scaffolds, which are intended for housing and industrial construction applications, mainly facade works. Similarly to the case of the two-person scaffolds described above, the platform is suspended by steel cables from booms laid on the roof and secured with a ballast or anchored in the roof.

The scaffold consists of 3–4.5 m long 1.5 m wide segments adding up to a desired length. A schematic of the sectional hanging scaffold is shown in Fig. 14.84.

Thanks to the platform's sectional design, hanging scaffolds of different shapes (angular and arched) can be formed, as shown in Fig. 14.85.

The vertical motion of the scaffolds is effected by an electric motor, cardan shafts, and winches.

The possible ways of anchoring the booms are shown in Fig. 14.86, and the specifications of sectional hanging scaffolds are listed in Table 14.14.

Mobile Hanging Scaffolds

Mobile hanging scaffolds can be moved horizontally on an industrial car without disassembling and reassembling the booms from which they are suspended.

A typical mobile hanging scaffold design is shown in Fig. 14.87.

The vertically moving work platform is suspended by steel cables from booms mounted in a swinging mode on an industrial car. Drives for traveling on a track laid on the roof of a building and for hoisting the platform are installed on the industrial car. The scaffold can be steered from both the platform and the roof. Scaffolds of this type usually have a hoisting capacity of about 300 kg and are capable of an elevation of 100 m.

14.6 Cranes

14.6.1 Mobile Cranes

Mobile cranes are intended for lifting and lowering loads and transferring them in the horizontal plane [14.23, 32]. Mobile cranes find wide application in the assembly of steel and reinforced concrete structures, repairs and materials handling. Their advantage is their mobility, high traveling speed, and quick setup on a construction site.

As regards their undercarriage, mobile cranes are divided into:

- Truck cranes
- Terrain-wheeled cranes
- Crawler cranes

Mobile cranes are typically truck-mounted. Cranes with a maximum lifting capacity of up to 20 t are usually mounted on mass-produced truck chassis, whereas high-capacity cranes are mounted on special undercarriages. The latter may have all their wheels driven and turnable, making them highly manoeuvrable and able to move over rough terrain.

Modern hydraulically driven cranes have replaced cranes with mechanical and pneumatic steering. A modern truck crane with electrohydraulic drives and steering is shown in Fig. 14.88.

The crane's base is a frame which, depending on the crane's design and hoisting capacity (maximum 20 t), may constitute a separate subassembly mounted on a typical truck chassis or be an integral part of a special truck chassis (high-capacity cranes). Mobile cranes are usually driven by internal-combustion engines, although some terrain-wheeled cranes and their working tools are driven by hybrid combustion–electric drives. In high-capacity cranes usually two driving motors are used: one for driving the vehicle and the other (situated in the slewing body) for driving the working tools.

The crane's frame incorporates four struts, with hydraulic lifts attached to their ends, extended by two hydraulic servos. For work the struts are protracted horizontally and then the whole crane is jacked up by the hydraulic lifts to such a height that the road wheels do not touch the ground. Some cranes can operate on their road wheels.

The slewing body with the operator's cabin is connected with the chassis frame through a crown bearing. All the working fittings are mounted on the crane body. The jib consists of a stationary section and two to six protractible sections. The jib's sections are usually protracted synchronously. The winch is mounted at the origin of the telescopic jib's stationary section. A significant number of cranes are equipped with an auxiliary jib which in its working mode can be attached to the telescopic jib's head in order to increase the crane's radius. In the traveling mode the auxiliary jib is attached to the side of the permanent telescopic jib. A counterweight is mounted at the rear of the slewing body.

The principal working gears and systems of telescopic-jib cranes are:

- *A jib-protracting gear* driven by two reversible hydraulic cylinders. As a rule, the jib's members are protracted synchronously.
- *A crane radius changer* driven by a reversible hydraulic cylinder. In high-capacity cranes sometimes two hydraulic cylinders are used for this purpose. The crane radius changer typically allows one to set the jib at an angle of 0–75°.
- *A slewing gear* driven by a hydraulic engine, a planetary gear, and a crown bearing with outer meshing. The slewing gear is equipped with an automatically controlled multiple disk brake.
- *A lifting gear* consisting of cable drum winch, a cable, pulleys, and a pulley block. The cable drum is driven by a hydraulic engine and a planetary gear. As a rule, the winch's pull force is much weaker than the crane's maximum lifting capacity. Hence it is necessary to use multistrand blocks to reduce the forces in the cable.

All the working gears and the gear that extends the struts are hydraulically driven. In truck-mounted low-capacity cranes equipped with one internal-combustion engine, the hydraulic oil tank and the hydraulic pumps are mounted in the crane's undercarriage. Hydraulic oil is fed into the body's working circuits through a hydraulic rotary joint. In cranes with two internal-combustion engines, the hydraulic pumps of the crane's working circuits are driven by the engine mounted in the slewing body.

The following working motions of the crane are controlled:

- Body slewing
- Change of jib length
- Inclination of jib
- Lowering and raising of the hook
- Traversing gear motions (for terrain crawler cranes and some wheeled cranes in which traveling with a load suspended from the hook is allowable)

These working motions are typically controlled from the operator's cabin. A system of control levers, which can limit the linking of the particular working motions, ensures proper control of the latter. The recommended systems and directions of motion of the control levers are shown in Fig. 14.89.

All cranes are equipped with the following safety devices:

- *A load limiter* – usually a microprocessor unit signalling that the hook block is reaching the rated load and disabling the crane's working motions when the rated load is exceeded. A signal indicating that the hook block is approaching the rated load is activated at 0.9–1.0 of the nominal load: an orange indicator light comes on in the control panel in the operator's cabin and an audible warning is produced. Overload is signalled by a red indicator light and disabling of the working motions, except for downward motion of the lifting gear. The signalling system is activated and motions are disabled when the crane block load is in the range of 1.0–1.1 of the rated load. The load limiter should be calibrated for a given hoisting capacity characteristic. Each load limiter should be equipped with an interlock for disabling the limiter in an emergency:
- *A block upper position limit switch* – disengages the winch's drive when the block finds itself at a certain distance from the jib's head.
- *Slewing gear limit switches* – are used in cranes that cannot turnaround completely, e.g., an allowed rotation angle of 270°. These protect the crane against the situation in which the slewing column and the jib reach an out-of-specification position relative to the chassis.
- *Cable unwinding limit switch* – protects against complete unwinding of the cable from the winch drum. The limiter is actuated when there are only a few coils of cable left on the drum.
- *Emergency jib retracting and lowering system* – enables the retracting and lowering of the jib to a safe position in the case of a failure of the hydraulic system.
- *Hydraulic system protections* – enables the automatic return of the piston rods of the struts', the radius changer's, and the lifting gear's cylinders.

The basic parameters characterizing the operating properties of mobile cranes are:

- Hoisting capacity
- Radius
- Hoisting height

The values of these parameters are usually presented in the form of diagrams representing hoisting

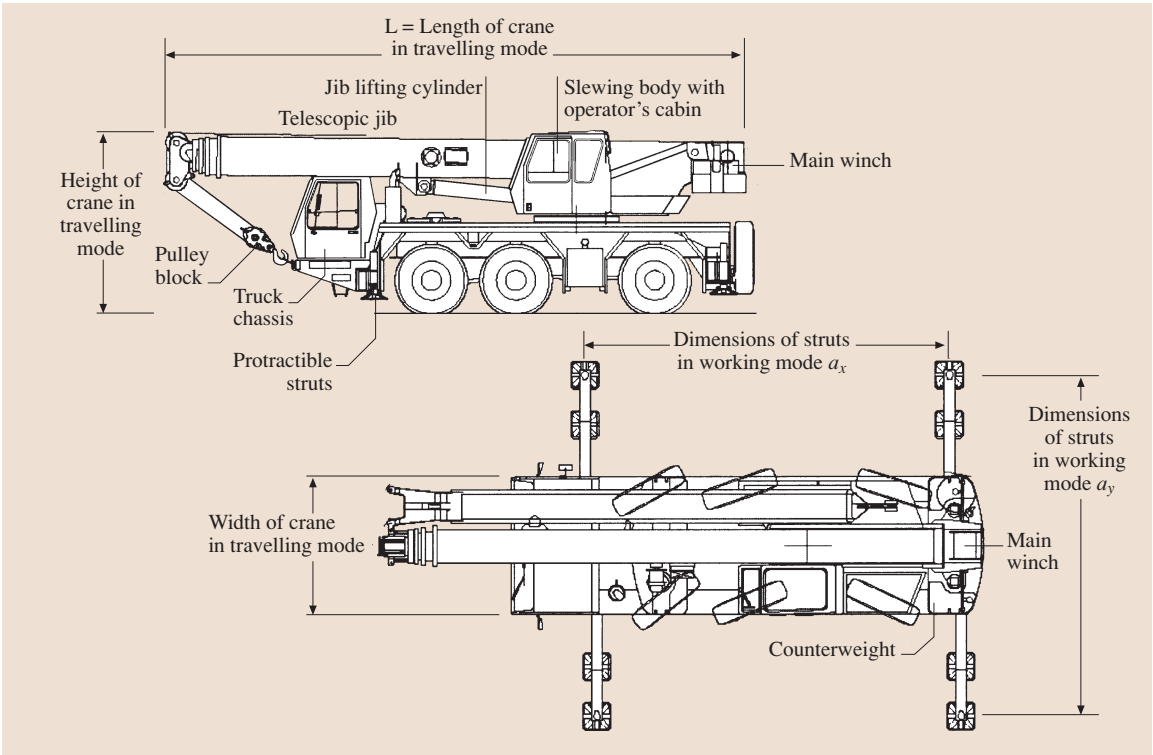


Fig. 14.88 Mobile crane mounted on a special truck chassis in traveling mode (a_x and a_y dimensions of struts in working mode)

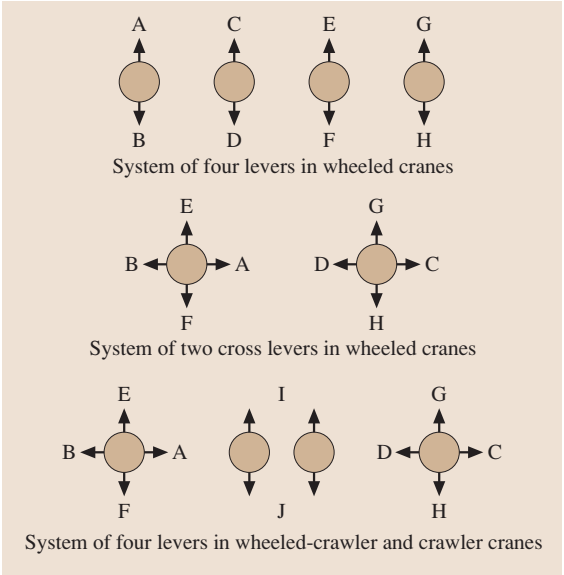


Fig. 14.89 Systems and directions of control lever motions in mobile cranes

capacity versus radius and jib length. Exemplary diagrams for truck-mounted mobile cranes are shown in Figs. 14.95–14.98.

Hoisting capacity diagrams specify calculated and experimental crane load values which take into account the crane's position stability and its structural strength.

Mobile cranes very considerably in their basic operating parameters. The hoisting capacity of currently manufactured cranes ranges from 5 to 550 t. Their basic specifications are listed in Table 14.16.

When selecting a crane for a given range of works one should consider the following:

Table 14.15 Symbols used in Fig. 14.89

A	–	turn right	G	–	lower hook
B	–	turn left	H	–	raise hook
C	–	protract jib	I	–	travel forwards
D	–	retract jib	J	–	travel backwards
E	–	increase radius	F	–	decrease radius

- The maximum hoisting capacity is given for the retracted main jib and a small radius of about 3 m and so is of little practical value. Hoisting capacity at larger radii decreases markedly.
- One should aim at operation at the smallest possible radii and hoisting heights so that a rational choice of a crane can be made and the crane's useful properties be effectively exploited. The use of large radii and hoisting heights should be technologically justified.
- One should take into account the travel of the crane to the working area and the bearing capacity of the ground on which the crane is to be set up. The bearing capacity of the ground should be appropriate for the anticipated loads acting on the struts.

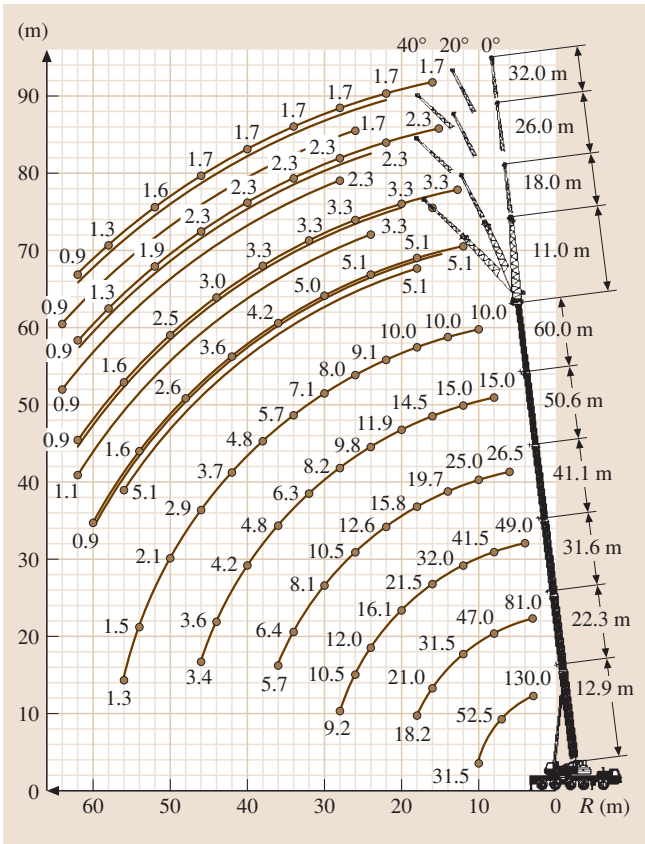


Fig. 14.91 Diagram representing hoisting capacity of crane with maximum capacity of 130 t. Note: The characteristic is determined for the main jib and the auxiliary jib. The *numbers* above the curves specify the allowable hoisting capacity for a given jib and hoisting height

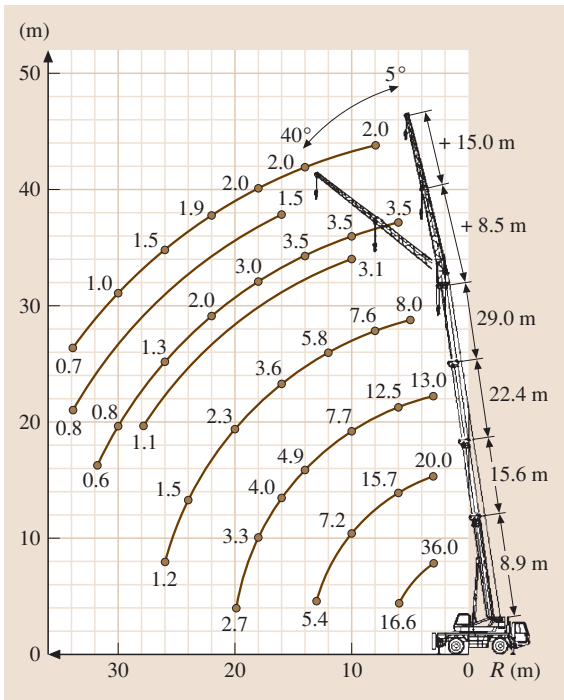


Fig. 14.90 Diagram representing hoisting capacity of crane with maximum capacity of 35 t. Note: The characteristic is determined for the main jib and the auxiliary jib. The *numbers* above the curves specify the allowable hoisting capacity for a given radius and hoisting height

14.6.2 Small Capacity Portable Cranes, Gantries, and Winches

A range of portable machines, based on winches and other accessories, for handling materials and transferring light equipment on construction sites has been developed.

This machinery includes:

- Scaffold cranes mounted on scaffolds (Fig. 14.92 items 5 and 6, and Fig. 14.95)
- Portable cranes (Fig. 14.92 item 4, and Figs. 14.96, 14.98) fixed to steel supports installed between floors, in window openings or on the roof (Fig. 14.97, basic parameters are shown in Table 14.17)
- Gantries mounted on the roof (Fig. 14.92 item 3), in an opening in the building's elevation (Fig. 14.92 item 2) or on a scaffold (Fig. 14.89 item 1)

The main component of the above machines is a universal winch that can work in tandem with various accessories.

Figure 14.92 shows the use of scaffold cranes, portable cranes, and small-capacity gantries during building erection.

Machinery of this type is intended for lifting and transferring loads of up to 200 kg to a height of 80 m. The design and technical specifications of these winches make them a highly effective means of vertical transport in construction work involving scaffolds as well as the assembly and disassembly of scaffolds.

Winches in scaffold cranes can be mounted in two ways:

- Outside the crane, to the lowest (from the ground) scaffold upright (Fig. 14.93)
- On the crane's boom (Fig. 14.92 item 5, and Fig. 14.94)

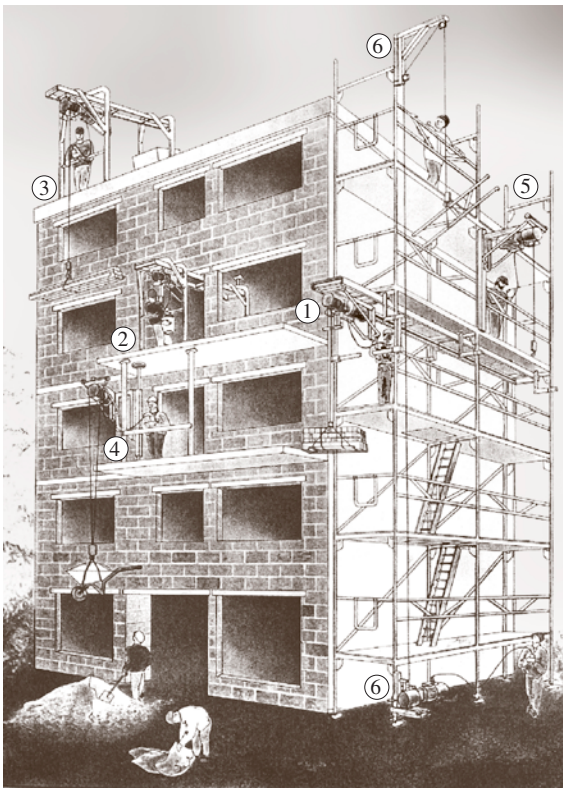


Fig. 14.92 Use of scaffold cranes, portable cranes, and small-capacity gantries during erection of building (items 1–6 are explained in the text)

In the case of winches mounted using the former method, a limit switch, functioning also as a load limiter and a block upper position switch, is incorporated into the winch's housing.

The way in which a winch is mounted onto the boom is shown in Fig. 14.94 and Fig. 14.92 item 6.

The working radius of the boom with a mounted winch can be changed by protruding the load-bearing tube. There is a series of holes in the inner tube for a blocking pin. The boom with the winch can be attached in a slewing mode to all kinds of support elements (Figs. 14.95–14.98).

The advantage of winches mountable on booms is their simple design and assembly owing to the

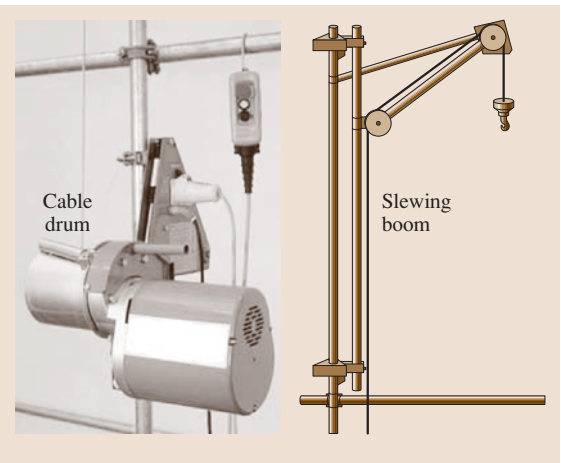


Fig. 14.93 Scaffold mountable winch

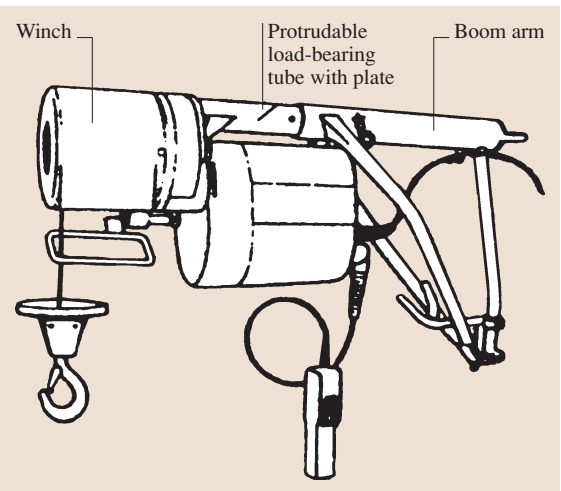


Fig. 14.94 Winch mounted on boom

Table 14.16 Specifications of selected mobile cranes with telescopic jib

Rated hoisting capacity	8	25	50	90	130	450
Minimum radius of main jib (m)	3	3	3	3	3	3
Maximum radius of main jib (m)	12	26	34	37	56	56
Jib's angle of inclination (°)	0–75	3–82	3–82	3–82	3–82	3–82
Cable winding speed (m/min)	42	130	120	125	120	130
Dimensions in traveling mode (mm)						
• Length	9050	10 225	11 020	12 795	14 980	19 622
• Width	2500	2500	2500	3000	2750	3000
• Height	3500	3450	3480	3795	3910	3990
Spacing between struts						
a_x (mm)	4450	6325	6625	8100	7800	8760
a_y (mm)	4136	6200	6200	7000	7500	8900
Maximum traveling speed	40	75	85	79	80	85

elimination of intermediate cable pulleys. Their disadvantage is the unfavorable weight and load distribution along the boom's end, resulting in the increase in the forces needed to slew the loaded boom and in heavier loading of the load-bearing structure.

The structure of winches with a hoisting capacity of 60–200 kg, employed in portable cranes and gantries, is shown in Fig. 14.99. The characteristic feature of such winches is the use of an electric motor with a built-in brake and the integration of all the units, i. e., the electric motor, the toothed gear, the drum, and the electric control system.

The drive units of modern winches commonly incorporate:

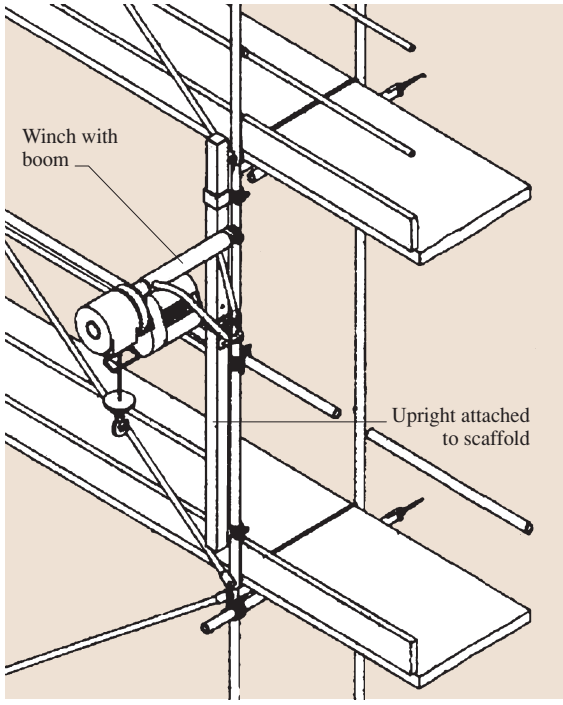


Fig. 14.95 Scaffold crane

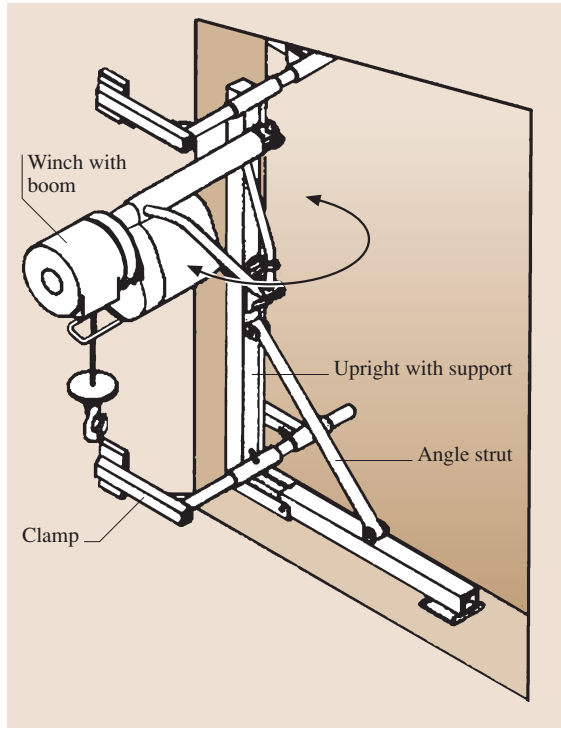


Fig. 14.96 Portable crane mounted in window opening

- Clutchless connection between the motor and the gear transmission – a gear wheel interacting with the transmission gear's toothed wheel is mounted in the rotor shaft's end.
- The cable drum is equipped with bearings internally whereby the transmission and the cable drum are compact.

Modern winches are intended for vertical transport for a wide range of construction works. Hence the range of handled construction materials is highly diverse as regards kind, shape, dimensions, and so on.

For this reason the manufacturers of light cranes offer a wide range of accessories for securing the load. Examples of accessories for handling different kinds of materials are shown in Fig. 14.100.

The use of such elements greatly increases work effectiveness and improves operational safety.

14.6.3 Tower Cranes

General Information

Tower cranes are commonly used in civil engineering for short-distance transport of loads and for erecting reinforced-concrete, steel, and masonry structures. Tower cranes owe their universality to the ease with which their structural-operating specifications (hoisting height, radius, and lifting capacity) can be adapted to the needs of construction sites. Such adaptation is possible thanks to the modular structure of tower cranes.

Tower cranes are classified according to the design of their basic units, e.g., the kind of slewing gear, the method of relocating and setting up, the jib design, and the tower design.

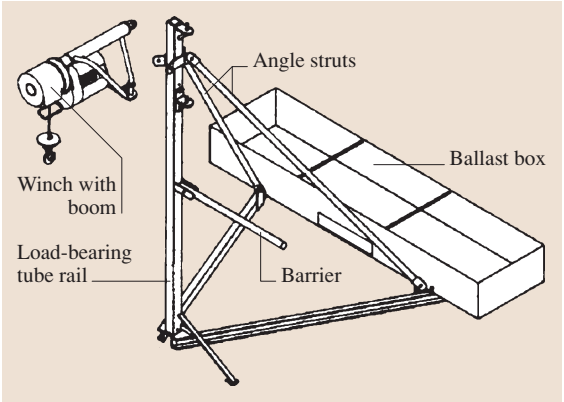


Fig. 14.97 Portable crane mounted on building's roof

Table 14.17 Technical specification of portable cranes

Lifting capacity	60–200 kg
Boom's radius (accessories)	About 1 m
Lifting height (max)	75–80 m
Lifting speed – first gear	15–20 m/min
Lifting speed – second gear	40–60 m/min
Accessories boom slewing	Manual
Power supply	230 V single-phase AC
Motor's power	0.3–1.1 kW
Control	Control panel
Drive unit mass	35–60 kg

According to the slewing gear design, two kinds of tower cranes: *high-* and *low-slewing* (with a non-slewing tower and a slewing tower) can be identified. The typical components of *high-slewing tower cranes* are: a stationary vertical tower resting on a base or a foundation anchor and a slewing part consisting of a rotary ring, a turret, a jib, and a counterjib. The typical components of *low-slewing tower cranes* are a slewing vertical tower resting on a base and a jib.

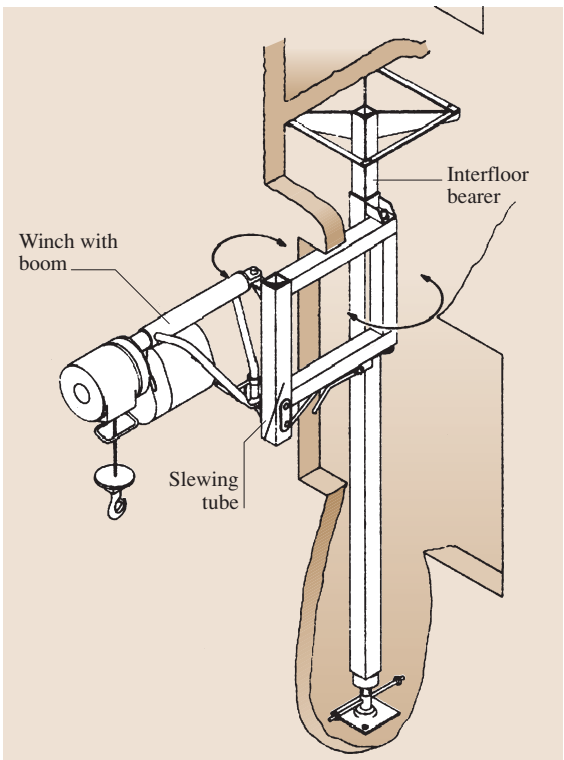


Fig. 14.98 Portable crane fixed to steel support

Each tower crane mounted on a base can be additionally equipped with a traversing gear to enable it to move on a straight or curvilinear track.

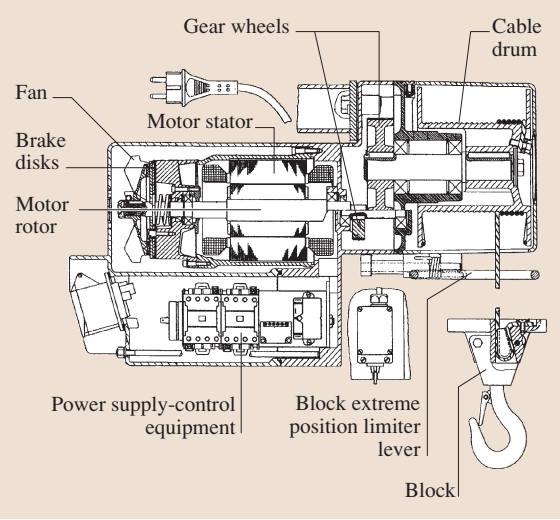


Fig. 14.99 Structure of winch used in portable and scaffold cranes and small-radius gantries

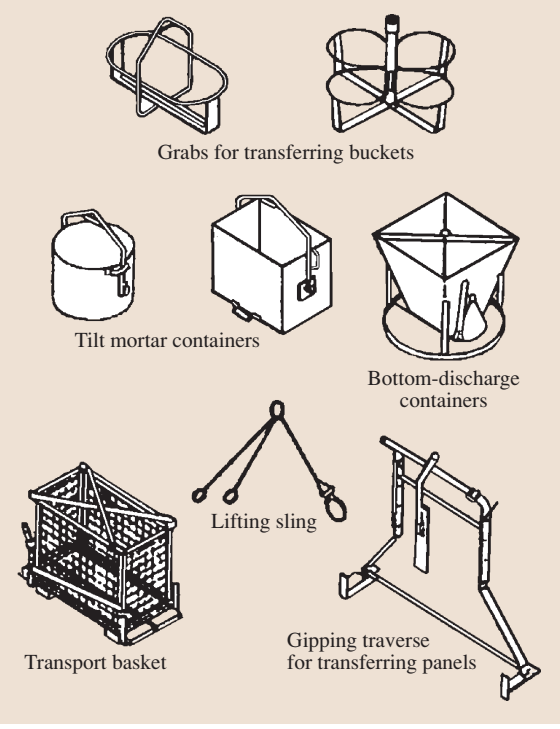


Fig. 14.100 Examples of accessories for handling materials by portable cranes, gantries, and winches

The tower and jib of modern cranes have a modular design. The tower's height and the jib's length can be adjusted by changing the number of tower or jib modules (sections). As a result, the basic operating specifications of tower cranes (hoisting height, radius, and hoisting capacity) vary greatly, not only within the particular kinds of cranes but also within a given crane model.

The main operating specifications of the most common type of cranes, i. e., *high-slewing cranes*, are as follows:

- Autonomous hoisting height: up to 100 m (at a 2.45 m × 24.5 m tower cross section)
- Maximum hoisting capacity: 50 t (4–16 t in housing construction)
- Radius: 20–80 m

Besides the above most common cranes, there are also cranes made to order, e.g., a crane with a hoisting capacity of 225 t, a radius of 80 m, and a hoisting height of 130 m. The crane is equipped with a computer system that controls the position of the jib counterweight to counterbalance the bending moments acting on the tower and it has a lift for transporting the operator to the control cabin.

The basic operating specifications of *low-slewing cranes,* as an option, are as follows:

- Maximum hoisting height: up to 50 m
- Maximum hoisting capacity: 2–8 t
- Radius: 10–50 m

Most of the currently manufactured *low-slewing trolley cranes* can operate with the jib raised at a certain angle (usually 8–30°).

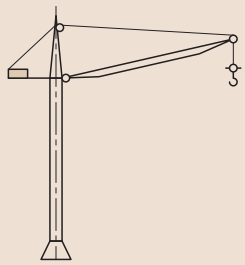
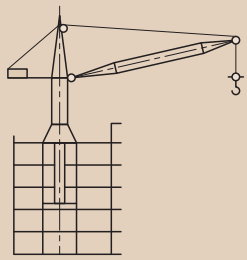
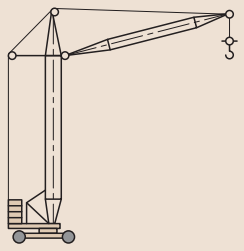
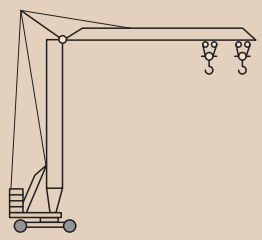
A classification of tower cranes according to their design features such as:

- Movability relative to the erected building structure
- Jib design
- Tower design

is detailed in Tables 14.18–14.20.

Cranes with a horizontal trolley jib are most commonly used in construction. Cranes with inclinable jibs find many fewer applications; they are used mainly on construction sites with restrictions on the crane's work radius because of the site's location and its organizational-legal conditions (e.g., in places where charges for the area over which the crane's jib passes are levied).

Table 14.18 Classification of tower cranes according to their movability

Type of crane	Description	Sketch
Stationary crane	Mounted on foundation anchor	
Floor crane (moving up with erected building structure)	Mounted on erected building structure and jacked up as successive storeys are built	
Track crane (moving on track)	Mounted on base equipped with traversing gear enabling traveling on track	
Mobile	Truck-mounted or wheeled, partially or fully folded during transport	

Cranes with Trolley Jib

As mentioned above, tower cranes with trolley jib are currently most commonly used on construction sites. A typical design of such a crane is shown in Fig. 14.101.

The crane's structure consists of basic units: a base, a tower, a slewing head, and a jib.

Base. Depending on the site's requirements, the crane's base can also be mounted on:

Table 14.19 Classification of tower cranes according to jib design

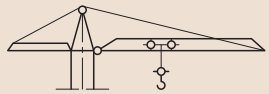
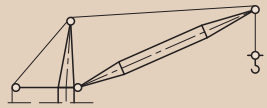
Type of crane	Description	Sketch
Crane with horizontal trolley jib	Jib operating in horizontal plane; radius is changed through movement of trolley along jib; jib can be attached to tower with or without tie rods. In some cranes, jib can be raised at angle of 8–30°	
Crane with inclinable jib (without trolley)	Jib can be inclined in vertical plane; radius is changed through inclination of jib	

Table 14.20 Classification of cranes according to tower design

Name of crane	Description
High-slewing tower cranes (cranes with nonslewing tower)	Crane's tower does not slew. Jib can slew thanks to slewing towerhead connected with tower via rim bearing
Low-slewing tower cranes (cranes with slewing tower)	Crane's tower slews. Tower is connected to crane's base via rim bearing

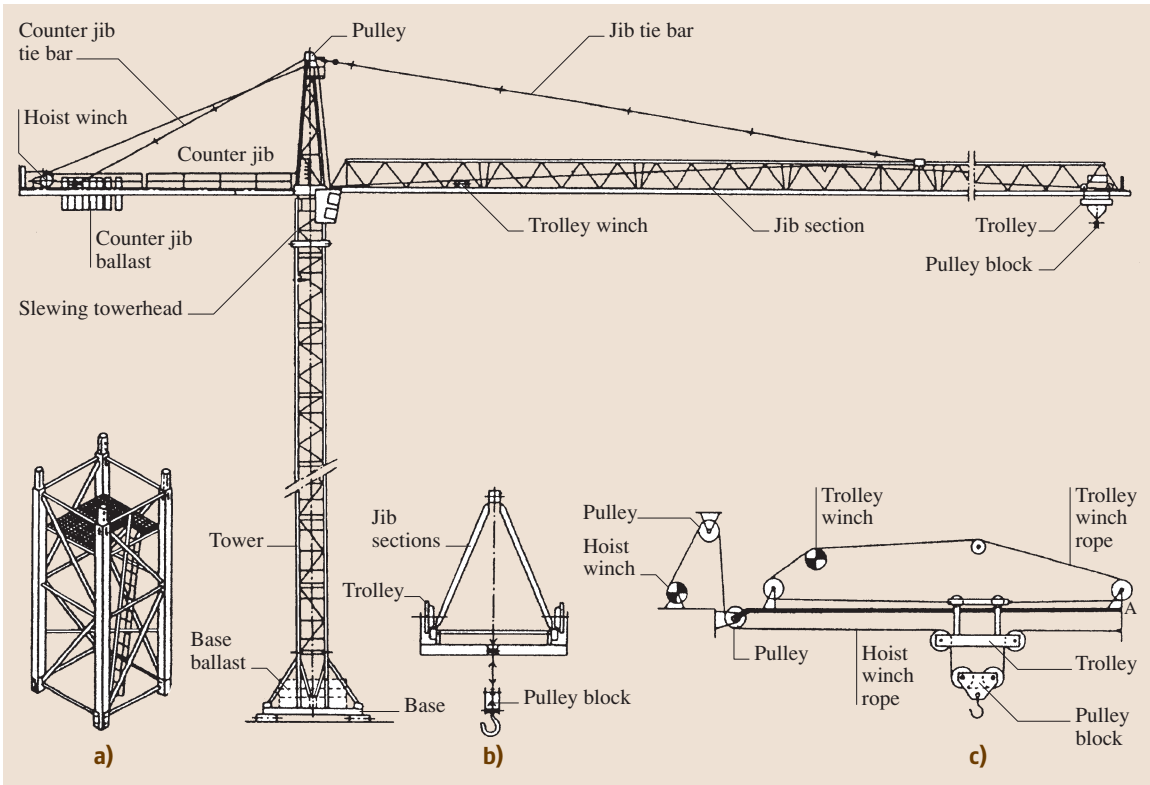


Fig. 14.101a–c Tower crane with trolley jib – main parts. **(a)** Tower's section. **(b)** Jib and trolley's cross section. **(c)** Kinematic diagram of trolley and pulley block winch hoist drive

- Supports provided with foundation plates (Fig. 14.102a)
- Trucks traveling on a rail-track (Fig. 14.102b)
- Supports on a concrete foundation (Fig. 14.102c)
- Feet embedded in foundation concrete (Fig.14.102d)

In most modern high-slewing cranes, the tower can be also mounted directly on the building structure being erected and goes up with it. For the assembly and disassembly of cranes three special frames are used. The tower is fixed by means of two frames to the erected building structure's members (usually a floor) (Fig. 14.103). A third frame is used to jack the crane up to a higher storey.

The *tower* is a metal structure usually in the form of a space truss with columns open or closed in cross section. The tower functions as a support for the jib which can be mounted on it at the appropriate height. In cranes with a nonslewing tower, a head is attached, via a rim bearing, to the tower's top. The tower is made as a welded construction or from sections joined by fasteners. Typically, 2.8–6 m-high (for the telescopic self-erecting model) sections square in their cross section or 6–12 m-high (for a crane set up using a truck crane) ones are used. The sections are fastened together with bolts or pins. In high-slewing cranes, a slewing head is mounted onto the stationary tower's top and a jib and a counterjib (with its ballast) are attached to it. The slewing head assembly usually incorporates a slewing gear drive.

In most cranes, the height of the tower can be increased without it being necessary to partially disassemble the crane. This is done using a telescopic cage by means of which the crane can be raised and additional tower sections can be inserted.

The *jib* is a space framework fixed to the tower (slewing-tower cranes) or a slewing towerhead (nonslewing tower cranes), making it possible to obtain a proper radius through the shift of the trolley. Jibs usually consist of several sections to facilitate their transport and assembly. The trolley moves on the jib's bottom flanges (Fig. 14.101b).

Pulley block hoisting gear. A hoist winch rope (Fig. 14.101c) passes from the hoist drum (Fig. 14.101c) via two pulleys (Fig. 14.101c) attached to the tower and then along the jib (Fig. 14.101) through the trolley (Fig. 14.101c) and the pulley block (Fig. 14.101c) to a securing point (Fig. 14.101c) at the jib's tip. As the rope winds on (unwinds from) the hoist drum, the pulley block is raised or lowered.

Trolley traversing gear. The trolley is shifted along the jib by a closed rope system (Fig. 14.101c) driven by a hoist winch (Fig. 14.101c).

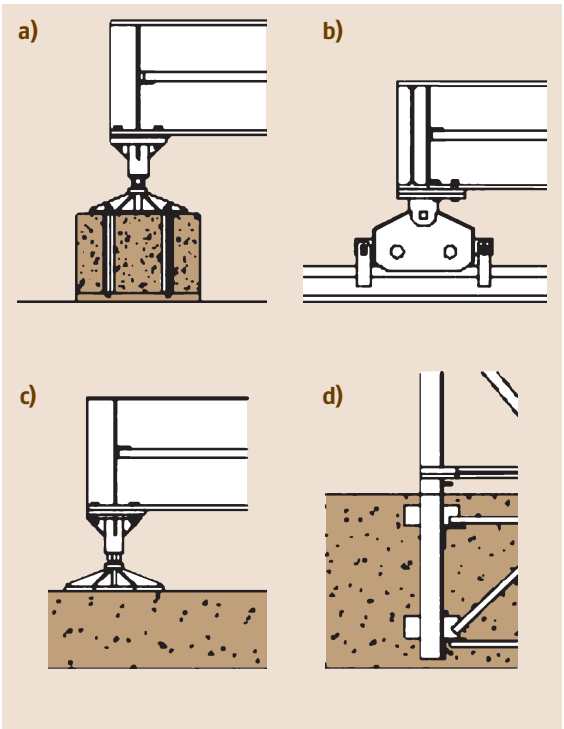


Fig. 14.102a–d Alternative ways of mounting tower crane with trolley jib **(a)** on supports provided with foundation plates; **(b)** on rail trucks (mobile version); **(c)** on supports and foundation plate; **(d)** on feet embedded in foundation concrete

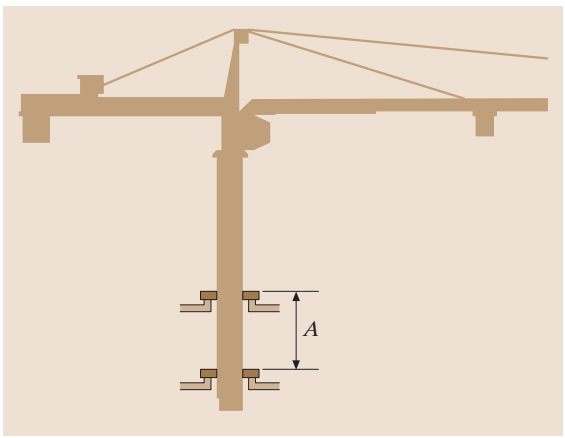


Fig. 14.103 Way of fixing tower to structural members of building under construction

Table 14.21 Technical operating specifications of typical 8–40 t hoisting capacity cranes

Parameter	Tower crane with maximum hoisting capacity of				
	8 t	10 t	12 t	25 t	40 t
Maximum hoisting height (m)	71	71	69	93	80
Maximum radius (m)	60	65	70	80	80
Maximum speed of hoisting two-suspender pulley block (m/min)	46 (4 t) 124 (1 t)	36 (5 t) 96 (1.25 t)	30 (6 t) 78 (1.5 t)	0–34 (10 t) 80 (2.5 t)	40 (20 t) 96 (2.5 t)
Minimum speed of hoisting two-suspender pulley block (m/min)	4	3	2.4	–	–
Trolley traversing speed (m/min)	0–76	0–50 (10 t) 0–100 (5 t) 0–120 (2.5 t)	0–50 (12 t) 0–100 (6 t) 0–120 (3.0 t)	0–47 (25 t) 0–63 (12.5 t) 0–76 (6.25 t)	0–33 (40 t) 0–50 (20 t) 0–100 (2.5 t)
Max. slewing speed of jib (rpm)	0.8	0.8	0.8	0.8	0.8
Possible ways of crane mounting	A, B, C, D	A, B, C, D	A, B, C, D	A, B, C, D	A, B, C
Tower section length (m)	3.3/5/10	3.3/5/10	3.3/5/10	3.3/5/10	3.3/5/10
Jib section length (m)	5/10	5/10	5/10	5/10	5/10
Load characteristic for max. jib length	Fig. 14.90 curve for 8 t crane	Fig. 14.90 curve for 10 t crane	Fig. 14.90 curve for 12 t crane	Fig. 14.91 curve for 25 t crane	Fig. 14.91 curve for 40 t crane

Legend:
A – base mounted on feet
B – base with suspension system enabling traveling on rail-track
C – tower mounted directly on feet embedded in foundation concrete
D – mounted on building structure under construction and jacked up as successive storeys are erected

A hoisting speed proper for the hoisting capacity is selected by a remote-controlled gear switch. The speed can be as high as 125 m/min. The maximum trolley traversing speed is 80 m/min. The pulley block hoisting gear and the trolley traversing gear are driven by squirrel-cage motors or slip-ring motors equipped with electromagnetic brakes. Hydrostatic drives or electric motors with controlled speed are also used.

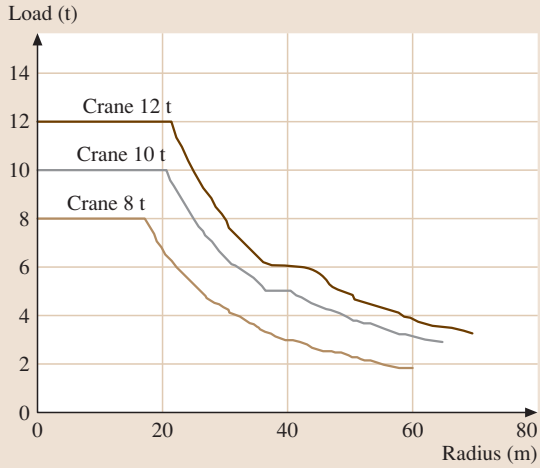


Fig. 14.104 Load–radius curves for 8 t, 10 t, and 12 t cranes

The crane's work motions are controlled from the operator's cabin using the controllers installed there, or they are radio-controlled. The operator's cabin can be equipped with a hoisting gear by means of which it can be hoisted to the appropriate height.

In the case of track-mounted (mobile) cranes, trackways made from traffic rails, (wooden or concrete) sleepers, ties, and stops, laid on a subgrade, are used. The track can be laid on a soil subgrade or a structural subgrade (a support structure, a building floor, a hard road surface). Usually railway rails are used for the tracks. The rails should be fixed to sleepers laid on a subgrade. The tracks should have protective groundings and lightning conductors. The tracks are equipped with the following protections:

- *Stationary and movable bumping blocks*: Stationary bumping blocks are usually situated at a distance of 1.5 m from the end of the track rails and movable bumping blocks at a distance of 1.2–1.5 m from the stationary ones.
- *Travel stops*: Devices switching off the crane's traversing gear when the gear's tripper runs into

a cam. Cams are mounted at the end of the crane track before the movable bumping blocks. The cams are positioned in parallel to the track so that the lever of the traversing gear tripper mounted on the crane's carriage can run into them.

After work, mobile cranes are secured to the trackway by means of clamps (fastening the carriage to the rails) preventing wind pressure from shifting the crane from its work position.

Tower cranes are equipped with the following protections:

- *Hoisting capacity limiter*: A device protecting the crane from lifting too heavy a load for a given radius. The protection acts in two stages. The reaching of the nominal hoisting capacity or moment is signalled acoustically and visually starting at 90% of the nominal values. The loading of the crane with a load amounting to 100–115% of the nominal hoisting capacity or with a moment amounting to 100–125% of the nominal moment results in the shut down of all the drives of all the gears except for the hoisting gear's lowering function. Each load limiter should have an interlock so that it can be switched off in an emergency situation.
- *Moment limiter*: Functions similarly to the hoisting capacity limiter.
- *Pulley block top hoisting position tripper*: Shuts down the hoisting winch's drive when the pulley block comes very close to the jib's trolley.
- *Rope unwinding tripper*: Prevents the rope from unwinding completely from the hoisting drum. The tripper is activated when only a few coils of rope are left on the drum.
- *Trolley extreme position tripper*: Stopping the traversing of the trolley in one direction in its extreme positions.
- *Slewing tripper*: Stops the slewing motion of the crane when the crane's angle of rotation in one direction exceeds 1.5 turn (540°) to prevent the machine's feeder cables and control cables from being damaged (this applies to cranes without a rotary joint enabling unlimited slewing).

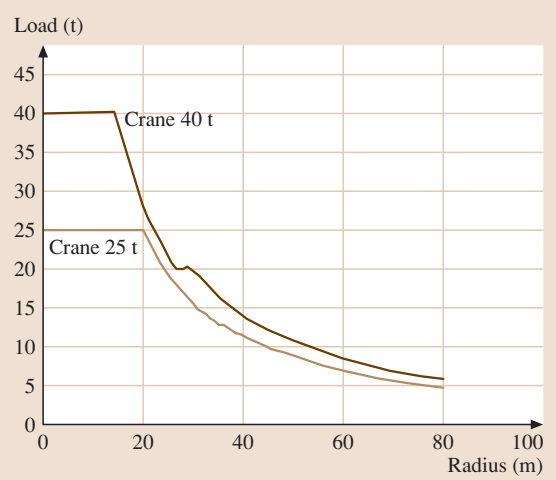


Fig. 14.105 Load–radius curves for 25 t and 40 t cranes

Since there is a wide range of tower cranes with a horizontal trolley jib, one can select the crane best suited to the construction site's conditions. It is common practice that several tower cranes with different hoisting elevations and capacities are so arranged around the structure under construction that they are complementary to one another. In such cases, particularly when the cranes' work areas overlap, it is extremely important that their simultaneous work is properly synchronized.

The basic structural-operating specifications of 8–40 t hoisting capacity cranes are shown in Table 14.21. For small radii the parameter which lim-

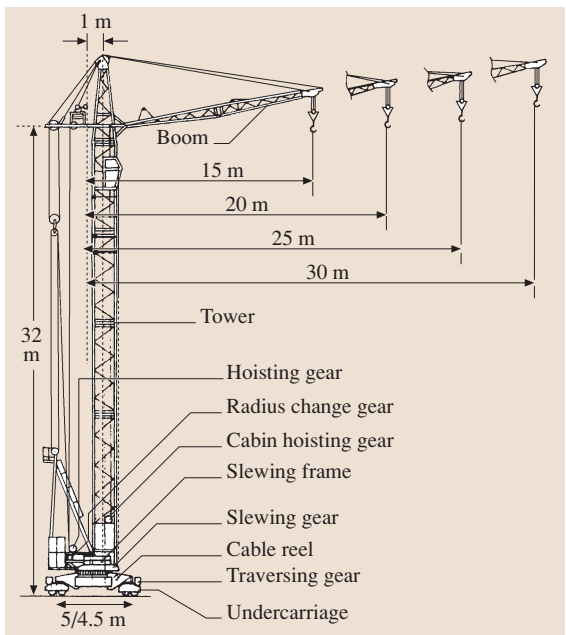


Fig. 14.106 Track tower crane with 10 t hoisting capacity with luffing boom

its the hoisting capacity of a tower crane is the strength of the load-bearing structure, while for larger radii the decisive parameter is the crane's stability. The crane's hoisting capacity characteristics specify the safe load values, calculated and experimentally determined taking into account the crane structure's stability and strength.

The load–radius curves for 8–40 t cranes are shown in Figs. 14.104 and 14.105.

Cranes with Luffing Boom

The structure of tower cranes with a luffing boom is described below using as an example a 10 t hoisting capacity track crane with a slewing tower (Fig. 14.106).

The crane's undercarriage consists of a welded box frame and cantilevers for mounting rail trucks. The undercarriage enables travel on straight and arched tracks. A rotary ring is attached to the undercarriage frame. A slewing frame with a hoisting gear, a radius changer, a slewing gear, and concrete ballast plates is mounted on the ring. The crane's tower and an installation cantilever are fixed to the slewing frame. The tower consists of segments, usually made from angle bars or steel pipes. An operator's cabin is attached to the tower, to which access is provided by a ladder secured to the tower. The crane's boom and a guy rope are attached to the tower's top segment. The boom is a welded lattice structure, quadrangular or triangular in cross section, made from high-strength steel sections. The boom consists of several sections (typically three ones) joined together by bolts. Cable pulleys, a hoisting gear tripper, and an overload tripper are mounted on the boom tip.

The crane's particular working motions are executed by the hoisting gear, the radius change gear, the tower slewing gear, and the traversing gear.

The hoisting gear consists of a motor, a coupling, a brake disk, a reduction gear, and a cable reel. Hoisting speed is changed by means of the reduction gear ratio lever. An additional drum for installing concrete counterbalance elements, is mounted on the hoisting gear's shaft.

The radius change gear consists of a hoisting winch and rigging. Drive is transmitted from the winch's motor via the flexible coupling working together with the brake and a thrustor to the reduction gear and the cable reel. The radius change reduction gear incorporates an additional friction–pawl brake. The cable system includes a stationary pulley block, a running pulley block, and guy-ropes. The slewing gear is secured to the slewing frame's front part. Drive is transmitted from the motor through the coupling with a brake disk and via a toothed gear to the rim bearing.

The traversing gear is made up of four rail trucks with double wheels, secured to the undercarriage's cantilevers. Drive from the motors is transmitted via couplings with brake disks, a reduction gear box, and an open-toothed gear to the wheels.

Similarly to cranes with a trolley jib, tower cranes with a luffing boom have the following protection devices:

- Load limiter
- Pulley block top hoisting position tripper
- Rope unwinding trippers
- Traversing limiters

The crane's working motions are controlled from the operator's cabin.

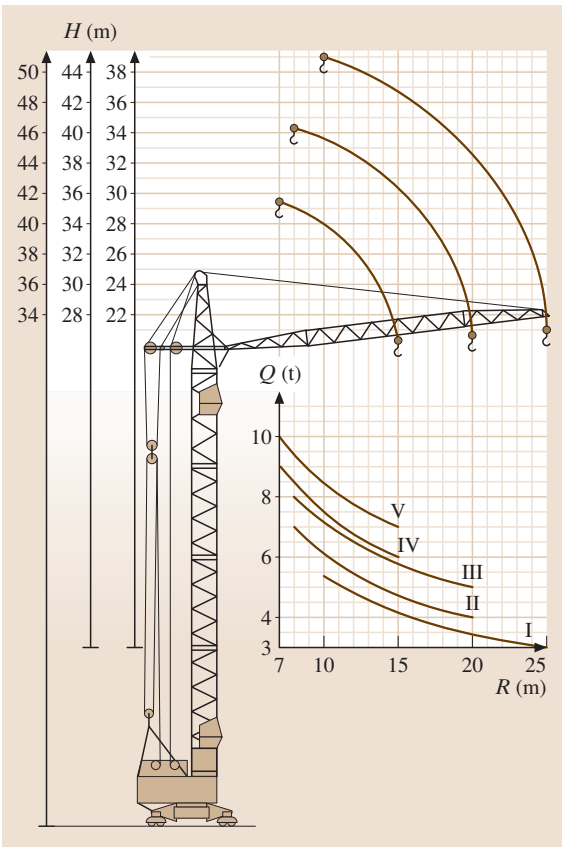


Fig. 14.107 Hoisting capacity as a function of radius and hoisting height for a 10 t-capacity crane with luffing boom

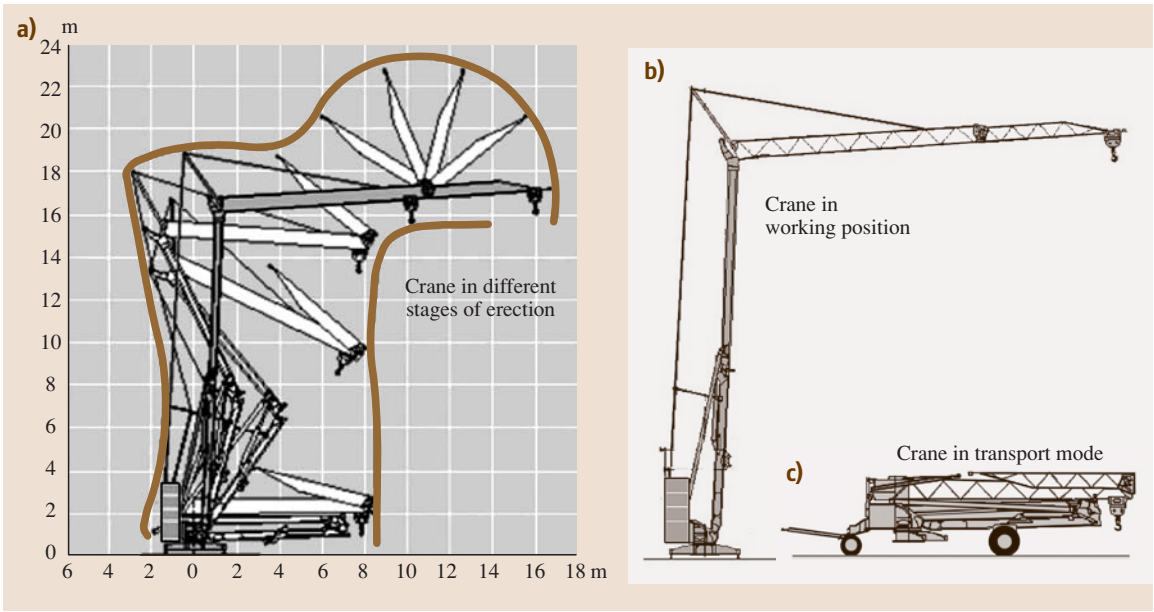


Fig. 14.108a–c Preparation of self-erecting crane for work: **(a)** crane in transport mode; **(b)** crane in different stages of erection; **(c)** crane in working position

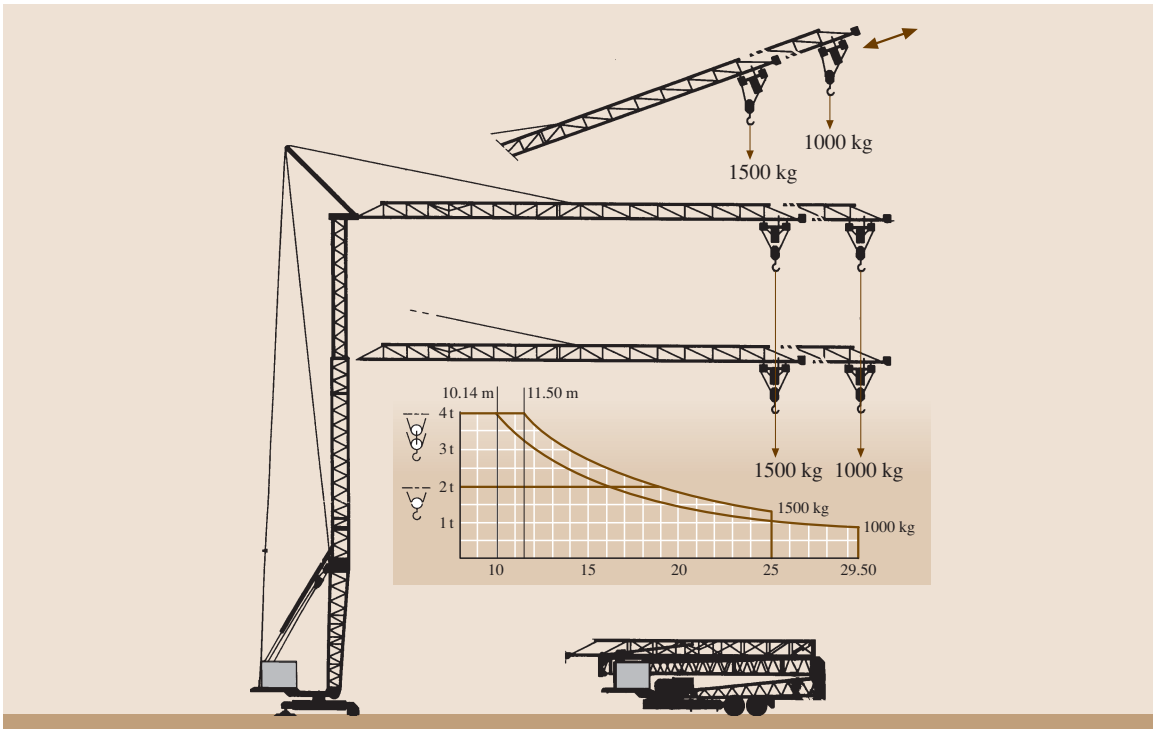


Fig. 14.109 Hoisting capacity versus radius for quick-assembling tower crane

A sample hoisting capacity–radius characteristic of a crane with a luffing boom is shown in Fig. 14.107.

Quick-Assembling (Self-Erecting) Cranes. Quick-assembling cranes form a separate class of tower cranes. Their characteristic feature is that they can be quickly assembled on a setup site. They do not require any other truck crane to be assembled, provided that the construction site has been properly prepared. The access road should be prepared such that the tractor towing the crane can reach the setup site and the latter should be practically at the same level as the access road. The setup site should be large enough for the vehicle with the ballasts to drive up close enough and for the crane to self-ballast (by means of its small auxiliary crane) and unfold. The crane is set up through the unfolding of the articulated mast and jib segments by its own drive units. During assembly the parts are connected by articulation but, once positioned, they are successively immobilized, forming a fixed load-bearing structure. The way in which self-assembling cranes are erected is illustrated in Fig. 14.108b, which shows four stages of assembly. Once the crane is properly set up, the jib's tip section is raised slightly. Then, using the crane's driving gears, its tower is put in a vertical position. Finally, the jib is completely unfolded.

Quick-erecting cranes are equipped with a trolley jib. They have a slewing tower which is secured to the base through a rim bearing.

An exemplary *hoisting capacity–radius* characteristic of a quick-erecting crane is shown in Fig. 14.109.

14.7 Equipment for Finishing Work

The aim of finishing work is to invest a structure with the design features and external and internal appearance. Finishing work includes:

- Roofing
- Outdoor (elevation) and indoor plastering
- Facing work
- Flooring
- Painting

The development of equipment for finishing work has been associated with the mechanization of the most labor-intensive and arduous activities. The first machines for finishing work were mortar pumps, followed by wood floor scrapers, parquet sanders, and mineral floor grinders. Later mechanical painting was introduced. Finishing work is carried out mainly on the construction site but efforts are made to move it to back-up facilities and transport ready-made elements to the site in order to increase work effectiveness. The range of finishing work is very wide in terms of both execution techniques and materials. The most commonly used equipment for finishing work is presented below.

14.7.1 Equipment for Roofwork

From the materials point of view the roofing used today can be divided into:

- Tar paper roofing mainly made from thermoweldable membrane
- Ceramic and stoneware tile roofing
- Bituminous shingle roofing
- Metal (zinc- and acrylic-coated steel sheet, rust-proof sheet, zinc sheet, titanium–zinc sheet, copper sheet, and other) roofing
- Polyvinyl chloride (PVC) panel and ethylene propylene diene monomer (EPDM) membrane roofing

Most roofwork is done using hand tools. For making *thermoweldable membrane roofing* devices equipped with liquefied petroleum gas (LPG) burners are used. There are two methods of making insulating coatings from thermoweldable membrane:

1. By means of a roofing machine and
2. Using only a set of burners

A roofing machine consists of the following units: a tar paper spreader, a battery of burners, a flexible gas hose, and an LPG cylinder. The burners' flames melt the layer of pitch on the tar paper and at the same time heat up the base. Under these conditions the tar paper is pressed against the base by a roller made of segments to ensure that the tar paper is pressed down along the entire width of the roll. On contact with the base the pitch cools quickly, forming a layer that bonds the tar paper to the base.

The set of burners includes a six-burner battery as well as a double burner and a single burner. The six-burner battery is secured to a steel frame equipped with two wheels. The single burner and the double burner are

used to lay tar paper strips that are narrower than the roll and in not easily accessible places.

All burner models are offered with a *sustaining flame* option, improving work safety. In recent years special burners for lap welding have been introduced.

Ceramic tile or bituminous shingle roofing is laid by hand.

Metal sheet roofing is hand-made from prepared elements. For the preparation of metal sheet roofing elements hand tools (shears, bending machines), or in the case of large-sized roofing work power-driven tools, are used. In the latter case, metal sheet is guillotined to the appropriate dimensions and formed using bending-flanging machines (Fig. 14.110). Roofing elements properly prepared in the site yard are transported to the construction site and built in.

Sometimes roofing is preceded by the laying of insulation materials. Insulation materials 20–27 mm thick are secured to concrete or steel bases by means of special nails and plastic elements (flanged bushes) driven in using cartridge-charged fixing tools. Nails and cartridge-charged fixing tools can also be used to secure PVC, EPDM, bituminous, and profile metal sheet roofing to bases.

14.7.2 Equipment for Plaster Work

Plastering machines are used for preparing, feeding, and rendering all kinds of plaster on the walls and ceilings of erected structures.

All of these activities can be performed by one machine, the so-called plastering unit, or a set of individual machines, consisting of a mixer, a mortar pump, distribution hoses, and a spraying nozzle. Hence in the machinery used for plaster work one can identify the following groups of equipment:

- Equipment for transporting prefabricated dry mortar from a silo to a plastering unit
- Mortar and plaster mix mixers
- Mortar and plaster mix pumps
- Plastering units

Equipment for transporting dry mortar from a silo to a plastering unit (Figs. 14.111 and 14.112) can be used to transport dry mortars with a bulk density of 0.5–2.0 Mg/m^3. The maximum distance over which the material can be fed is usually 200 m and depends on the compressor's capacity, the material's properties, and the lie of the transport conduits. The equipment operates as follows (Fig. 14.111): when the plastering unit's reservoir fill-up signalling system signals that there is no mortar, the silo closing valve opens and the pressure vessel fills up. Once it is filled, the valve closes and the compressor starts blowing compressed air through the aeration fabric into the vessel. Liquefied mortar is pumped through a hose from the vessel to the plastering

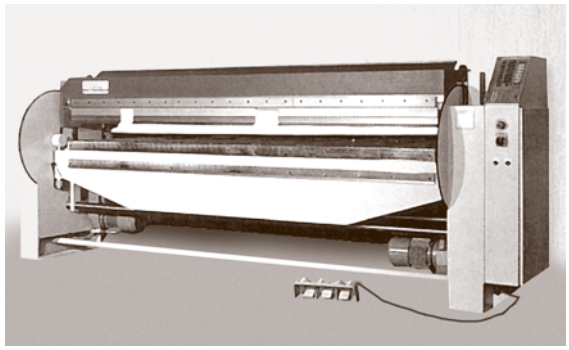


Fig. 14.110 Bending–flanging machine with roofing metal sheet profile programming

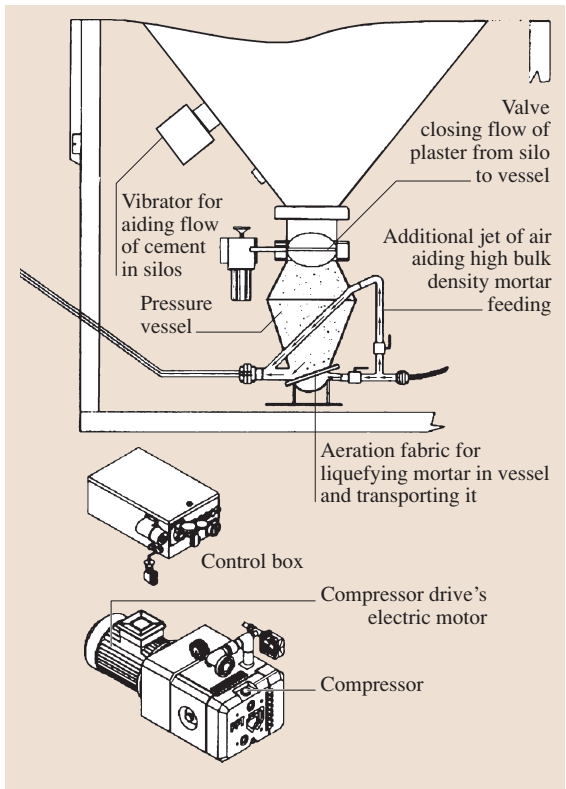


Fig. 14.111 Equipment for transporting dry mortar from silo to plastering unit

unit. If mortar with a high bulk density is to be pumped, the valve should be opened in order to aid the flow of mortar with an additional jet of air.

Portable equipment for transporting dry mortar, working in tandem with a plastering unit, is shown in Fig. 14.112. Besides a vibrator, an aeration unit is introduced in order to ensure proper flow of cement from the silo to the reservoir. The cement in the reservoir is liquefied by means of blow-in nozzles. The plastering unit's mortar reservoir is equipped with a lid with a fill-up signalling gauge and an air filter.

Dry mortar feeding systems can be made as mobile (equipped with a driving axle) or portable.

Mixers for Mortars and Plaster Mixes

Mixers for mortars and plaster mixes form a highly diverse class of plastering machines in terms of their size and principle of operation. The following kinds of machines are distinguished:

- Electrically driven, hand-operated mixers
- Continuous-type mixers
- Batch-type mixers

The machines are used for preparing masonry mortars, plasters, self-leveling mixtures, and so on.

Electrically driven, hand-operated mixers work by one or two electrically driven agitators that are manually introduced into a container to mix the contents.

Continuous-type mixers can be equipped with an open hopper (Fig. 14.113) or a dry mortar bin working with a dry mortar transport system (Fig. 14.112).

A diagram of a continuous-type mixer is shown in Fig. 14.113. In this mixer dry components are fed into the hopper and transferred by raking-out paddles and the feeding screw into the mixing unit, where water is added. The rate of flow of water into the mixer is controlled by a valve and a flowmeter.

Batch-type mixers perform the same function as continuous-type mixers, except that their operation is periodical and consists of the charging of components, their mixing, and discharging in succession. Their design is similar to that of concrete mixers. Pan-type mixers and paddle mixers are used. The most popular among the pan-type mixers are turbo mixers, planetary mixers, and turbo-planetary mixers. Among paddle mixers the most popular are mixers with a single helical agitator. Paddle mixers are discharged by tilting them or by opening a segment of the bottom.

It should be mentioned that also rotor mixers can be used to prepare gypsum mortars and mortars containing plastics. The components are mixed by setting them in rotary motion by means of a rotor.

Most mixers, both continuous and batch type, can work in tandem with mortar pumps.

Mortar pumps, when equipped with a mixer and a spraying gun, function as plastering units. The first mortar pumps were membrane pumps. Their maximum mortar pumping pressure was limited by the strength of

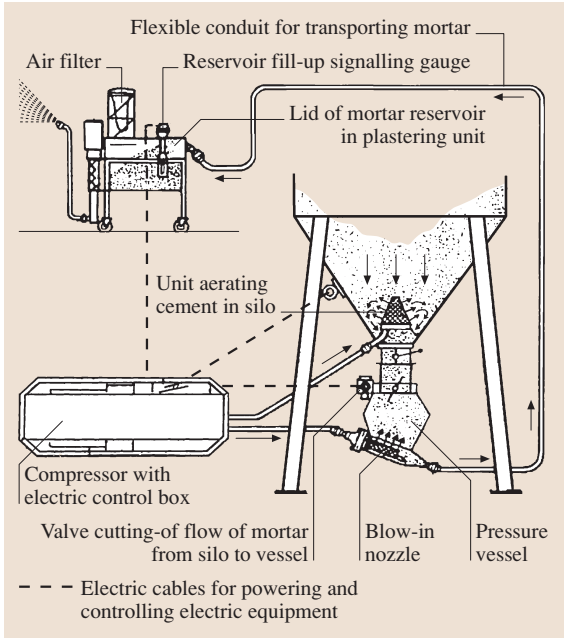


Fig. 14.112 Plastering unit working in tandem with equipment for transporting dry mortar

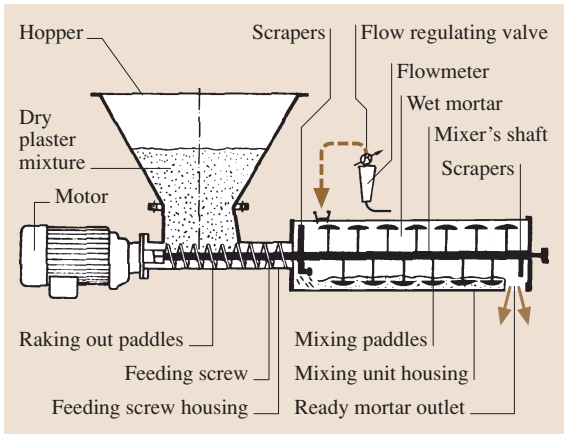


Fig. 14.113 Continuous-type mortar mixer with open hopper

the rubber membranes. As a result, the pump's lift was limited to about 20 m under average conditions. The desire to extend the reach of plastering and the required mortar forcing pressure has led to the development of piston pumps. The piston pumps currently produced by leading manufacturers can pump even heavy mortars to an elevation of 100 m at a forcing pressure of 6.0 MPa. The plaster feeding distances achieved depend on the forcing pressure, the mortar's composition, and the delivery rate. The delivery rate of piston pumps does not usually exceed 3.0 m^3/h.

A mortar pump design is shown in Fig. 14.114. The pump works such that, as the piston moves to the right, mortar is sucked through the suction valve into the working chamber. As the piston moves to the left the pressure of the mortar closes the suction valve and opens pressure valve. As the piston moves again to the right, mortar is sucked in again while the mortar on the piston's right side is forced into a pipeline. The pump makes it possible to minimize pressure fluctuations and therefore to maintain constant mortar spraying parameters and increase the fatigue strength of the mortar pipeline's flexible hoses. The travel of the pistons occurs as a result of the rotation of the cam pushing the roller during, respectively, the delivery stroke and the suction stroke.

The operation of a popular two-piston mortar pump with a cam drive is illustrated in Fig. 14.115. The function of the compensating piston is to equalize the mortar forcing pressure during the suction stroke of working piston. The suction stroke of the working piston is aided by a spring and the return stroke of the compensating piston results from the mortar pressure.

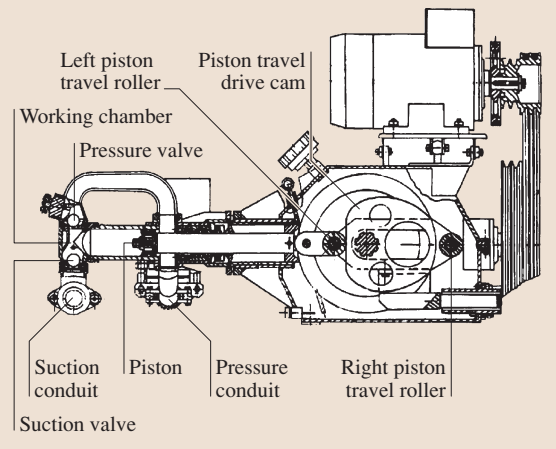


Fig. 14.114 Mortar pump with bilateral action piston

Besides piston pumps, screw pumps form another class of pumps. They are used for applying thin finishing coats of gypsum mortars, lime–gypsum mortars, and mortars with plastics. In recent years they have also been used for conventional mortars, competing with piston pumps. The principle of operation of the screw pump (Fig. 14.116) is to force mortar in the closed spaces between the helical surfaces of the rotor and sta-

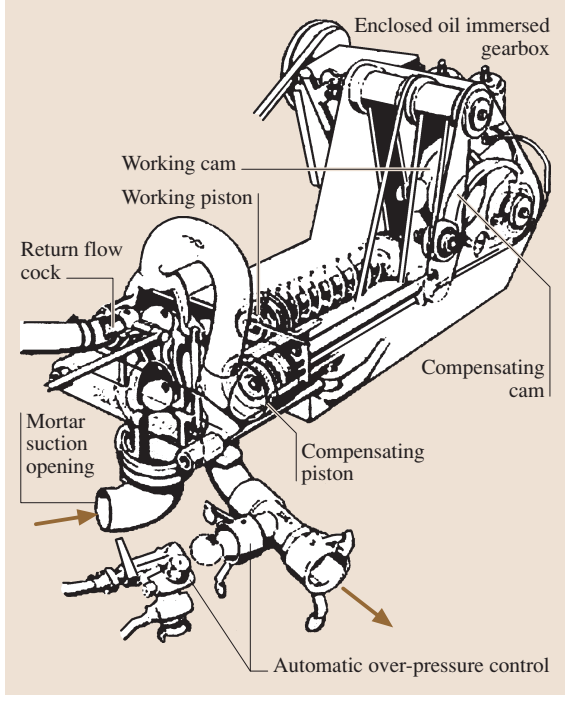


Fig. 14.115 Diagram of two-piston mortar pump with mechanical cam drive for the pistons

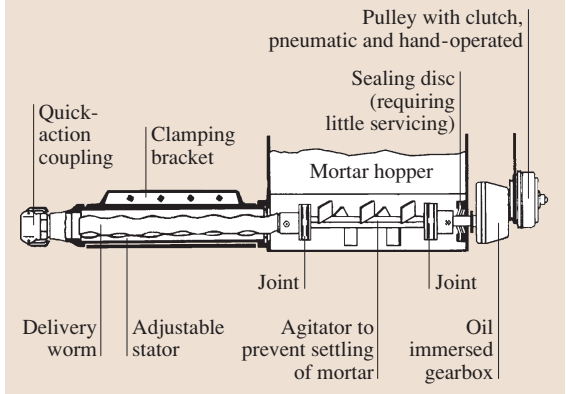


Fig. 14.116 Worm pump with adjustable stator

tor. The pump's stator is made of rubber that is resistant to compression and chemically aggressive liquids. The screw pump's efficiency decreases with operating time as a result of abrasive wear of the rotor and the stator. As the delivery of the pump decreases so does the forcing pressure. For the pump shown in Fig. 14.116 a decrease in pump delivery is counteracted by compensating the clearance between the rotor and the stator by means of a clamping bracket. Mortar is fed into the screw pump usually from the mortar hopper equipped with an agitator, which prevents the settlement of solid particles and the segregation of the material.

Plastering units are machines that mechanize the preparation and rendering of mortar. In the case of conventional cement–lime mortars, a plastering unit performs the functions:

- Charging of mortar components
- Mixing
- Straining
- Transport to rendering site
- Rendering on walls and ceilings

The introduction of plaster-like mixes and putties limited the activities connected with the preparation of mortar to only its transport or to mixing a dry mixture with water.

Depending on the plasters being made, plastering units can be divided into three classes that differ in their structural aspects:

- Plastering units for traditional lime and cement–lime mortars
- Plastering units for gypsum mortar
- Plastering units for plaster-like coats (PVC mortars)

A plastering unit for making traditional plasters from lime, cement–lime, and cement mortars is shown in Fig. 14.117.

In some plastering units, similarly to in concrete mixers, charging buckets are used in order to facilitate the loading of the components into the mixers. Plastering units are equipped with a remote control system for controlling the operation of the pump. If the spraying gun's air valve is closed, the pump's drive is automatically switched off and mortar feeding stops. The opening of the air valve results in the switching on of the pump's drive. As dry plaster mixes have become increasingly popular, plastering units for traditional mortars increasingly often feature screw pumps besides piston pumps.

A plastering unit for feeding and rendering mortars from ready-made dry plaster mixes is shown in Fig. 14.118. After refitting the pump and changing the mortar feeding hose, the plastering unit can also be used for self-leveling floor compounds.

A characteristic feature of the plastering unit shown in Fig. 14.118 is the reduced size of the mixer, whose function is performed by a mixing chamber with an agitator in the form of helical segments.

The rate of delivery of plastering units for traditional mortars is usually up to 3 m^3/h.

For feeding traditional plaster mortars, hoses that are 52–58 mm in diameter with tip elements 32–36 mm in diameter are usually used. For feeding and spraying special media, pressure hoses with increased strength are used.

Plastering units for traditional mortars can be adapted for spraying mixes to protect steel structures against fire, self-leveling mixtures, and similar materials.

Plastering units for gypsum mortars perform the functions of feeding, mixing, pumping, and pneumatic spraying of mortar onto a surface to be plastered. Their design is very similar to that of plastering units for dry plaster mixes. The structure of a plastering unit for gypsum mortars is shown in Fig. 14.119.

The mortar mixing chamber can operate in two positions: vertical and inclined. A vertical mixing chamber position is used for spraying ready-made gypsum plaster mixes. If the mixing chamber is inclined, a stroke pump can be used. For mounting and dismounting of the pump the mixing chamber is placed in a horizontal position.

A characteristic feature of plastering units for gypsum mortars is the common drive for the agitator and the pump, located in series.

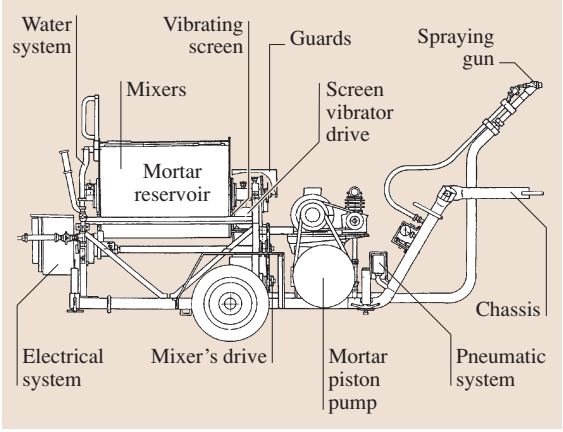


Fig. 14.117 Plastering unit for conventional plasters

In modern plastering units mainly screw pumps are used because of the following advantages:

- Pressure stability (practically no pressure fluctuations)
- Rate of flow can be adjusted by:
  - Changing the speed of rotation of the screw pump's rotor by controlling the speed of the driving motor
  - Adjusting the clamping pressure of the metal mantle on the stator (if the pump has such an option)

The rate of delivery can also be changed by replacing the pump with a unit with the desired flow rate.

Another useful feature of the screw pump is the possibility of changing the direction of flow by reverse pump rotation in the case of blockage of the pressure hose.

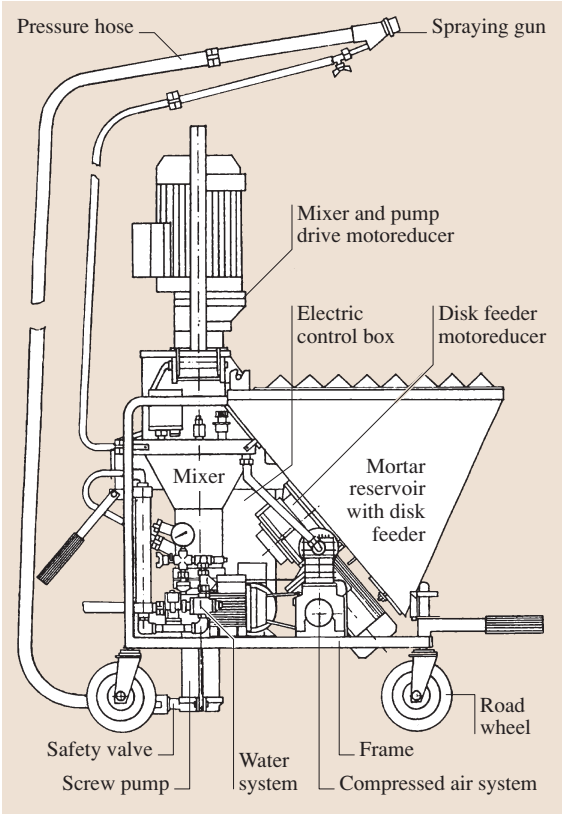


Fig. 14.118 Plastering unit for preparing and rendering mortars made from ready-made dry plaster mixes

The gypsum mortar flowing out of the spraying gun is torn apart by an axially introduced jet of compressed air and ejected from the nozzle at an accelerated speed. This ensures better adhesion of the rendered layer to the base. The shape of the jet can be modeled by adjusting the depth of insertion the compressed air nozzle into the spraying gun and by attaching spray-gun tips with different outlet diameters. Hoses with an inside diameter of 25 mm (sometimes 19 and 32–36 mm) are usually used for feeding gypsum mortars.

Electric motors with a power of about 5 kW are used to drive plastering units for gypsum mortars. They make it possible to feed mortar for a distance of 20–40 m at a working pressure of up to 4.0 MPa.

Plastering units for plaster-like coats (mortars containing plastics) are compact devices weighing a few tens of kilograms. They mix, pump, and render thin plaster coating mortars, adhesives, paints with fillers, dense insulating fluids, gypsum mortars, and so on.

The structure of a plastering unit for plaster-like coats is shown in Fig. 14.120. The reservoir can be charged with dry mortar, adding water subsequently, or with a mortar–water mixture. In order to ensure easy access for the spraying gun (particularly in window

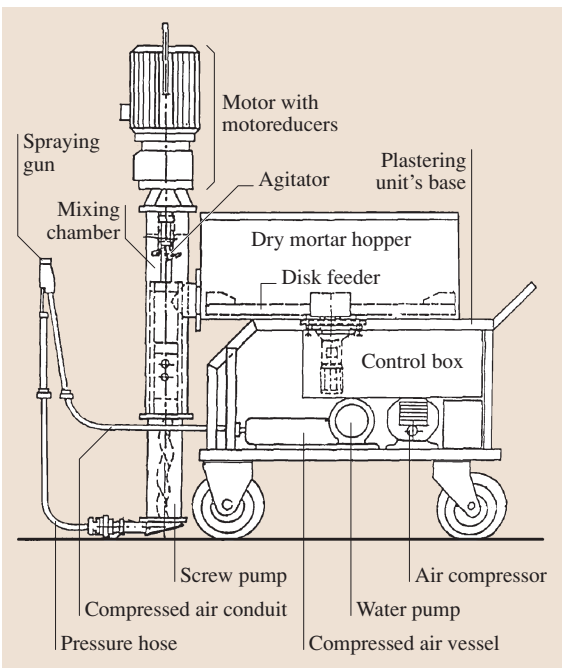


Fig. 14.119 Plastering unit for gypsum mortars

and door openings) screw pumps with a delivery of 0.3–0.7 m^3/h are used.

Hoses that are 25 mm (occasionally 19 mm) in diameter are used for feeding plaster mix. The simplest plastering units of this type can work in tandem with an external general-use air compressor. Plastering units are offered with an electric or diesel drive.

For operational safety reasons pressure conduits in all kinds of plastering units are protected by safety devices against an excessive increase in mortar or plaster mix pressure. Currently in most plastering units electromagnetic protection in the form of *pressure cut-offs* are used. Plastering units for conventional mortars, equipped with a piston pump, usually have double protection systems.

14.7.3 Equipment for Facing Work

Facing work consists of finishing surfaces by fixing decorative layers of different materials to them. The range of facing work is very wide. The most commonly used facing materials are: ceramic tiles, stoneware, and natural and artificial stone tiles.

The basic machine for facing work is a cutting-off machine. Such machines can be used not only for cutting the aforementioned materials, but also to cut concrete blocks, reinforced concrete, bricks, roof tiles, and so on [14.39].

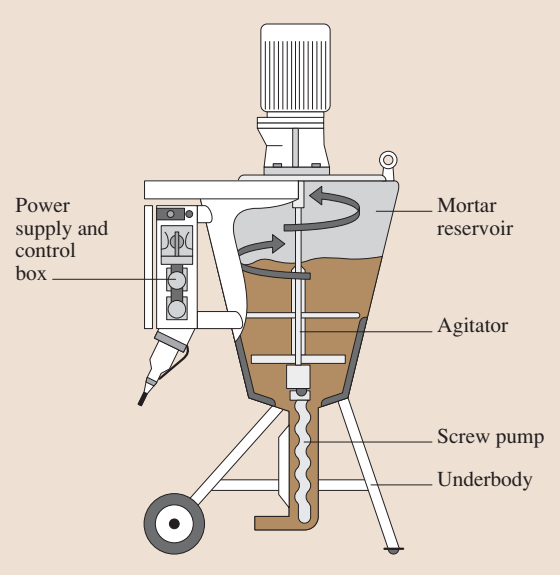


Fig. 14.120 Plastering unit for mortars containing plastics

Four types of cutting-off machines can be identified (Fig. 14.121):

- Type 1: A machine with a movable table having a fixed (permanently or by means of clamps) or swinging moveable cutting head (tiltable or not), which is located over the table
- Type 2: A machine with a fixed table having a horizontal moving cutting head and, if applicable, vertically adjustable and tiltable cutting head located over the table
- Type 3: A machine with a fixed table having a vertically moving cutting head
- Type 4: A machine with a fixed or movable and/or inclinable table having a fixed cutting head, intended for use only with continuous-rim tools having a maximum diameter of 250 mm (the motor being located under the table)

Cutting-off machines are used for cutting along straight lines, in planes inclined at various angles. The working tool is usually a solid or segmental disk with a diamond glued to it along its circumference, forming a ring. The disk is cooled with water. For less precise cutting (e.g., of bricks) disks with a diamond and metal can be used. The cutting disk is usually driven directly by a motor or sometimes via a belt transmission.

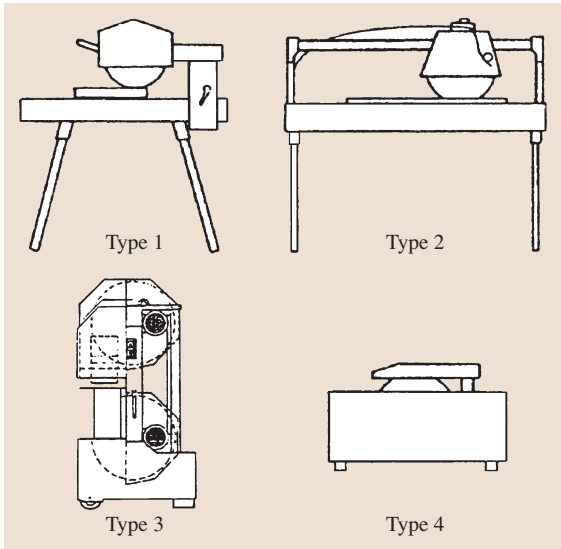


Fig. 14.121 Sketches of different types of cutting-off machines

The cutting disk is cooled and the plate of material being cut is wetted by immersing part of the disk in a tank with water, or by feeding water, flowing down gravitationally from a tank located over the disk or forced by a water pump, to the cutting disk. Wet cutting prolongs the life of the disk and eliminates dust emission.

The structure of a cutting-off machine is shown in Fig. 14.122.

The power of driving motors in cutting-off machines ranges from 200 W to 3 kW.

The group of machines used for facing work also includes a mixer for preparing adhesives and mortars. A mixer for preparing 40 l of mortar by mixing dry components with water is shown in Fig. 14.123. In order to fill bin one should tilt, by means of the handle, the mixing unit into the upper position and lock it with bolt. The handle is used to relocate the mixer.

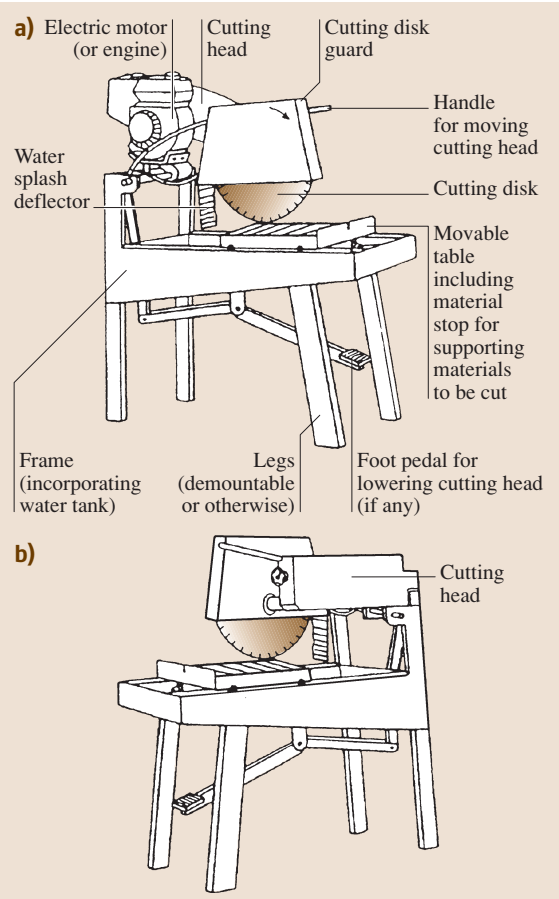


Fig. 14.122a,b Examples of cutting-off machines: **(a)** cutting-off machine with internal combustion engine; **(b)** cutting-off machine with electric motor;

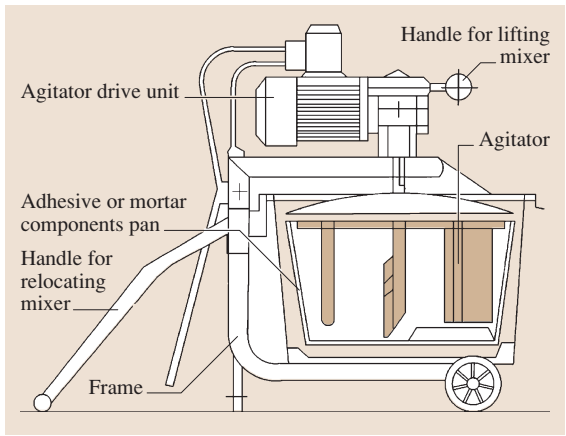


Fig. 14.123 Mixers for adhesives and mortars used in facing work

14.7.4 Floor Work

Floors are elements of buildings consisting of several layers and made almost entirely on the construction site.

Depending on the floor materials, different machines and equipment are used. The machines most frequently used for floor work are:

- Pneumatic feeders for dense mortars (feeder of fresh concrete mix and mortar)
- Vibrating beams (described in Sect. 14.3)
- Floating machines for concrete (described in Sect. 14.3.8)
- Grinders for stone and mineral floors
- Sanding–polishing machines for wooden floors (parquets)

A pneumatic feeder for dense mortars is used for mixing the components of cement mortar and delivering the latter to the placement site. A scheme of such a feeder is shown in Fig. 14.124. The vessel filled with the appropriate amount of water can be charged with mortar dry components manually or by means of a charging bucket. Compressed air supplied to the vessel and dosing unit forces mortar out of the vessel, but in the dosing unit the stream of mortar is separated by compressed air. As a result, in the pressure hose it already has the form of a series of small portions separated by spaces filled with compressed air. This transport system ensures stable flow of mortar into the hose's tip placed on

a tripod. Depending on their size, machines operating on this principle have a mortar capacity of 2–6 m^3/h and can feed mortar to an elevation of 50 m.

Electric motors or self-ignition combustion engines with power of up to 30 kW are used for driving pneumatic feeders for dense mortars. The most common pressure hose diameters are 63 and 68 mm.

Grinders for stone and mineral materials are being increasingly less frequently for finishing work in construction because of a change in the materials used. Commonly used until recently, terrazzo – which required grinding – has been replaced by lining in the form of stone and ceramic tiles with a ready-made smooth surface not requiring any machining after laying.

Wood floor sanders and sander–polishers form a numerous group of machines and find application in the construction of new structures and in renovation. The following types of machines can be identified:

- Drum (single- and double-drum) sanders
- Disk sanders
- Oscillating sanders

Drum sanders are intended for rough and finishing sanding of wooden floors. Sanders with one and two working drums can be distinguished. In the latter case, one drum performs sanding while the other stretches the abrasive belt. Both drums have a horizontal axis of rotation. The sander's working element is an abrasive (sandpaper or abrasive cloth). The sanding drum's working width is up to 250 mm and is limited by the necessity for the drum to exert appropriate linear pressures (about 20 N/cm) on the surface. The design of a single-drum sander is shown in Fig. 14.125.

A drum sander may weigh as much as 90 kg and the power of the driving motor can reach 3.5 kW at a drum rotational speed of about 2300 rpm.

Electric motors are exclusively used to drive sanders because the latter are intended mainly for work indoors and the possibility of ignition of the wood dust collected in dust bags has to be eliminated.

The described drum sander cannot sand the floor under heaters. For this purpose *disk sanders* (Fig. 14.126) are used. The working tool in this machine is an abrasive disk mounted at an angle of about 3° relative to the base. Sandpaper is Velcro-fastened to the disk or clamped with a nut. Because of the disk's inclination, its working base–contact surface amounts to about a third of the disk's surface area.

Power is transmitted to the disk by a V-belt or a cogbelt. The driving motor's power does not exceed

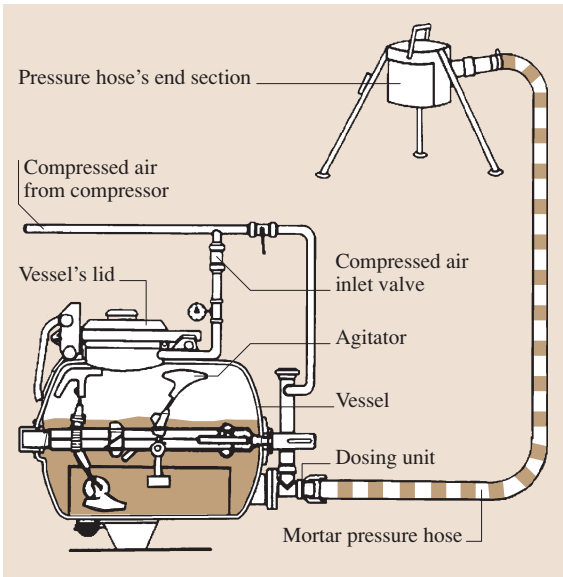


Fig. 14.124 Scheme of pneumatic feeder for dense mortars

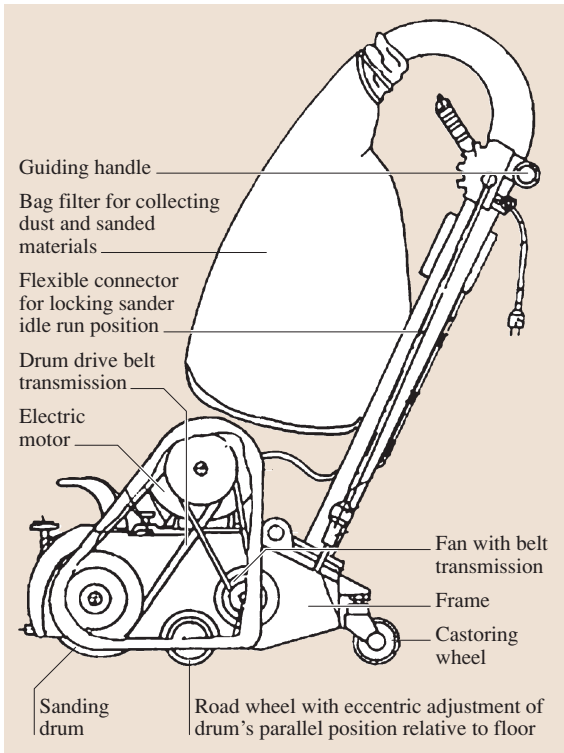


Fig. 14.125 Single-drum sander for parquet floors

2.0 kW at a rotational speed of about 4300 rpm. In currently manufactured disk sanders, disks about 180 mm in diameter are used.

In some sander designs, lights are fixed to the sander's body to illuminate the sanded surface at a small angle. This makes it easier for the operator to notice any surface irregularities and to obtain the required quality of sanded surface.

To sand corners of rooms and places not easily accessible to drum sanders and disk sanders *oscillating sanders and belt sanders* are used. These are typical mechanized devices with an electric drive.

Besides the above basic machines for floor work many small mechanical devices are used [14.40]. Their selection depends on the floor materials. This group includes:

- Industrial vacuum cleaners, thermal and mechanical cutters for polystyrene foam, sweepers, floor-washers, floor polish spreaders, and floor polishers
- For wooden parquet floors: hand-guided circular saws, circular saws with a table, and circular saw–fretsaw machines
- For laying plastic flooring: welding cord groove millers, circular tools, and welders

14.7.5 Equipment for Painting Work

Paint coatings are a very common way of finishing both building interiors and elevations. Paint coatings are applied to the surfaces of walls, ceilings, window and door openings, pipes, and heaters. Commonly used materials are emulsion, synthetic, and oil paints. The use of water (limewash and size color) paints is on the decline. In small rooms, paint work is as a rule done by hand. Painting units are used to apply paint coats in large rooms.

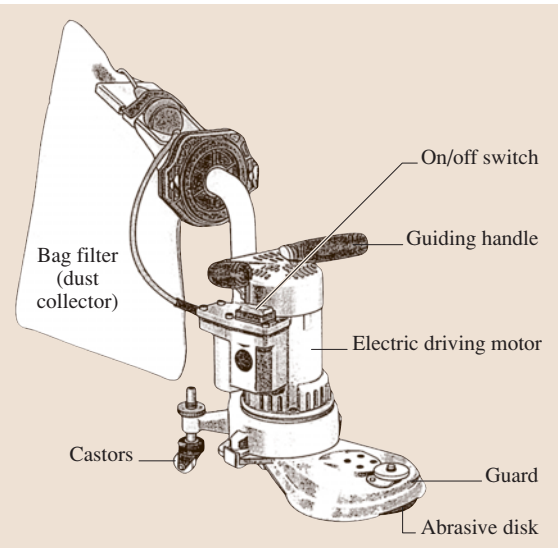


Fig. 14.126 Disk sander for parquet floors

Depending on the compressed air or paint pressure, painting units can be divided into three groups:

- Low-pressure (up to 0.55 MPa) painting units
- Medium-pressure (up to 2.5 MPa) painting units
- High-pressure (up to 25 MPa) painting units

Besides the above, electrostatic painting units are also distinguished.

Low-pressure painting units usually consist of a spraying gun, a paint reservoir, and an air compressor.

Spraying guns equipped with a paint reservoir are general-use devices. Paint reservoirs can be attached to a spraying gun from the top or bottom. There are also designs in which a jet of compressed air from a compressor is used to spray paint. The spraying parameters are set by means of valves, which control the paint and compressed air flows.

In medium-pressure painting units, similarly to in high-pressure painting units, paint is pumped without the action of air and so they are often called *airless*. Currently they are superseded by high-pressure painting units.

High-pressure painting units are used in construction for painting large surfaces and on roads to paint railings and roadway signs. The basic working unit is a piston pump which ensures a pressure as high as 25 MPa in the pressure hoses. The pumps and the paint forcing system's components are made of abrasion-resistent materials (tungsten carbide). High-pressure hoses, 12–16 mm in diameter, are usually employed. The hoses are protected against electric charge accumulation – typically by a conductor connected to earth.

Spraying gun nozzles ensure the rapid discharge of atomized paint in the form of a jet whose shape depends on the nozzle's shape (e.g., conical or flat triangular). The kind of nozzle and the spraying pressure are selected to suit the kind of work, e.g., the painting of large surfaces or the precision painting of stripes. The operation of modern high-pressure units is simplified as far as possible.

High-pressure painting units can be used for applying water-based paints, solvent paints, latex coats, acrylic coats, and so on. These machines are powered

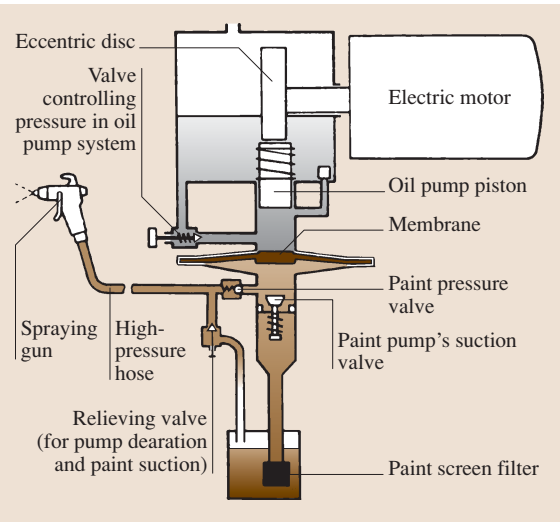


Fig. 14.127 Scheme illustrating the operation of a high-pressure (airless) painting unit

by electric motors, pneumatic engines or combustion engines, with a power as high as 4.1 kW.

A high-pressure painting unit is shown schematically in Fig. 14.127. The piston causes the pulsation of a membrane and the suction of paint from the reservoir through a filter and its forcing into the hose and spraying gun. A valve in the oil pump system is used to set the spraying pressure.

A typical high-pressure (airless) painting unit design is shown in Fig. 14.128. Paint is sucked in from a reservoir by the paint pump and forced (at a maximum pressure of 17.5 MPa) into the spraying gun with a nozzle 0.43 mm in diameter. The rate of painting is about 4 m^2/min.

It should be noted that in paint work, besides the application of paint coatings, preparation of surfaces and paints plays a vital role. Industrial vacuum cleaners, sliding grinders, and corner grinders are used for preparing the base [14.40].

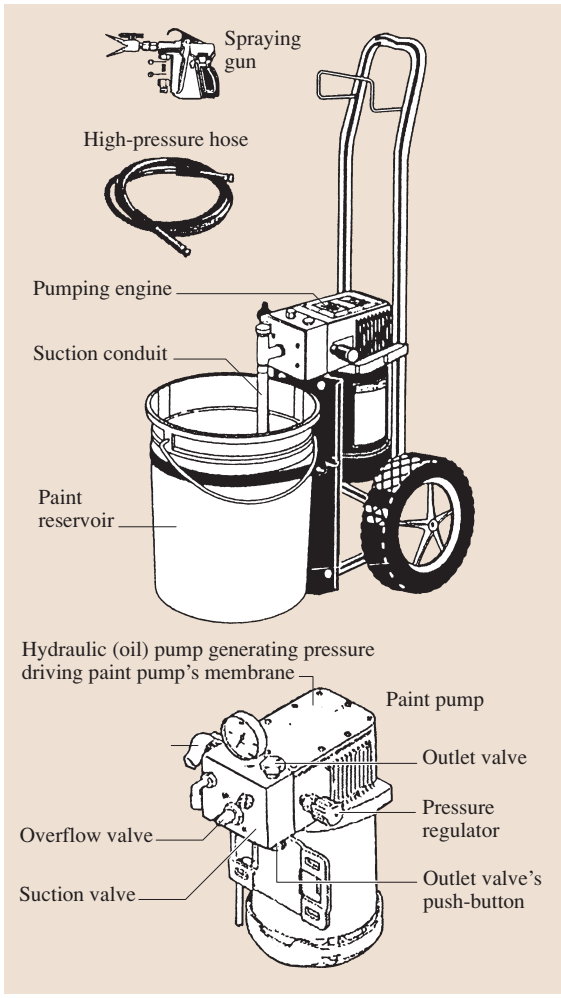


Fig. 14.128 High-pressure (airless) painting unit

Paints are mixed by means of hand-operated and mechanical mixers and strained using screens to remove impurities.

14.8 Automation and Robotics in Construction

Even today many people employed in construction associate the use of robots with the manufacturing industry, mainly the automotive industry, where they are typically employed to spray paint and spot weld the car body. However, considerable advances have been made in robotics applications in construction.

Robots are used not only because of their technical–economic advantages; they also result in improved work quality, increased output, reduced task realization time, reduced labor costs, and improved operational safety. The latter applies especially to work in conditions that are hazardous and detrimental to health, such as sewer

inspection, demolition work, underwater work, work in radioactive environments, earthwork on slopes, and work at heights. In many cases, it is for safety reasons that robots must be employed. A survey has shown that about 300 devices are currently available in the robot and automated equipment market.

The rapid development of robotics for construction applications in Japan began in earnest in the 1980s. Its direct cause was the shortage of labor in construction. Japan's business community turned to the government for permission to import workers from abroad. Because of the government's refusal, construction companies were forced to invest in research and development of robotics for automation in the execution of construction work. Another motivation was the difficulty with conventional performance of tasks that were deemed hazardous or arduous for human workers to perform. Waseda University's Systems Science Institute in Tokyo pioneered Japanese concepts of implementing industrial robotics on construction sites. Leading engineering construction firms such as Shimizu, Obayashi, Kajima, Taisei, Fujita, and others produced their own construction robot prototypes for various applications on construction job sites. These applications included single-purpose construction robotics as well as entire robotic systems for the performance of high-rise building construction.

In the USA the Robotics and Field Sensing Committee of the American Society of Civil Engineers engaged in coordinating research in automation and robotics for construction. The centers of active in this field included Carnegie Mellon University (Pittsburgh, PA), the University of Texas at Austin, Purdue University (West Lafayette, IN), the University of Southern California (Los Angeles), and North Carolina State University (Raleigh). Carnegie Mellon was a US pioneer in these developments, having started independent research and development of construction robotics in the early 1980s, shortly after the early developments of construction robotic concepts in Japan.

Also in the 1980s, construction automation and robotics concepts were researched in the former Soviet Union. These included systems and hardware developments at the Moscow Civil Engineering Institute and at the Central Laboratory for Construction and Heavy Manipulators [14.41–44]. Subsequent or parallel developments took place in France (*Centre Scientifique et Technique du Batiment*), Germany (the Fraunhofer Institute for Production Automation, Technical University of Karlsruhe, and the Technical University of Munich), the United Kingdom (Lancaster University, City University in London, and the former Bristol Polytechnic), and Spain (the Carlos III University in Leganés, Madrid).

As a result of the development of automation and robotics in construction in the last three decades this domain, integrating the achievements of robot technology, information technologies (IT), and design for construction (DfC), has acquired the status of a scientific discipline. Twenty years ago the International Association for Automation and Robotics in Construction (IAARC) was formed. This association organizes annual International Symposia on Automation and Robotics in Construction (ISARC). The 21st ISARC was held in Jeju (South Korea) in September 2004, and the 22nd ISARC took place in Ferrara, Italy in September 2005. Papers published in the symposium proceedings represent the latest achievements in this field.

In automated construction equipment two kinds of devices can be distinguished: teleoperation manipulators, referred to as construction manipulators, and construction robots. Construction manipulators are remote-controlled by the operator while construction robots are autonomous computer-controlled devices. The robot's software allows it to perform variable tasks within its application range.

In the development of automated equipment, four stages corresponding to the particular generations of this equipment can be distinguished [14.45].

The first generation, which can be called *automated construction devices*, was developed by outfitting existing construction equipment with electronic sensors and digital control. The principles underlying the development of the first-generation robots are still used in the automation of many types of construction machines. Expensive construction machines are equipped with sensors and computer control. The latter includes a data-processing unit and feedback control. Such adaptations are used in excavators, cranes, pile-driving equipment, transport to the horizontal transport of output, and in concrete mix transport and placement.

The second generation is associated with the application of manipulators to such construction works as laying reinforcement, building walls out of building blocks, floating concrete surfaces, and laying tiles. Manipulators are newly designed devices but still controlled by operators.

The third generation includes autonomous robots with no operator involved in their control. They need an operator only to prepare them for work and sporadically during work. They find application in many kinds

of construction work such as trenching, masonry work on construction sites, and in precast concrete plants for wall elements production, the assembly of steel structures, the transport of materials to the place where they are to be built in, the spraying of fireproofing insulation onto steel structures, and painting. Moreover, they are used for testing building structures and elements, e.g., sewers, the adhesion of tiles to tall buildings' elevations, and testing the quality of welded joints in steel structures.

Fourth-generation robots are designed for specific structures, taking into account the materials to be used. They are employed in automated building construction systems (ABCS). These robots are designed to be an integral part of a new construction methods which are adapted to the use of construction robots, known as design for robotic construction (DfRC).

All four generations of automated construction equipment are used in construction, and transition from one to another has been evolutionary. A major feature of this process is that the role of the operator is reduced, or completely replaced, by computer control.

Automated construction equipment used for particular kinds of construction work is described below.

14.8.1 Automation of Earthwork

Because of its significant share in the total building production, automation of earthworks deserves special attention [14.46]. However, some factors in earthwork make its robotization difficult, including:

- Variable forces of the ground-working tool interaction, due to the variability of the physical parameters of soil and material nonhomogeneity
- Variable height of the terrain
- Occurrence of buried objects such as electric cables, pipelines, etc.
- The possibility that machines working close to the edge of excavation edges may overturn
- The potential hazard to workers who find themselves within the machine's range of operation

For these reasons, remote-controlled machines and machines with a robot control system effecting the working tool motions controlled by computer-driven programs are usually employed in earthwork automation instead of true robots.

Such machines are used in work environments with radioactive and chemical contamination, in pile driving, underground work, tunneling, deep point-excavation, diaphragm excavation, and pneumatic caisson work.

The automation of earthmoving machines is proceeding in three directions:

- Use of remote control in machines
- Adaptation of machine control systems for automated execution of specific kinds of work
- Development of autonomous robots

The first direction (remote control of earthmoving machines) is the most common commercially available solution. On the basis of a three-dimensional image the operator controls the operation of machines and the loading of the excavated material. Remote control can be combined with a semirobotic control system when quality workmanship is required.

The second direction (adaptation of the machine's control system to make its fittings perform a specific task) includes, for example, the steering of the excavator bucket so that it moves along a predetermined trajectory [14.47–50], the control of the dozer blade to ensure that a smooth, leveled surface is obtained, and the control of drilling attachment mounted on an excavator diaphragm wall excavation.

The third direction is the development of autonomous third-generation robots for earthwork. The robot's design should be adapted to adverse service conditions and hazards.

Example Applications

The research and development work on the automation of earthmoving machinery is primarily concentrated on single-bucket excavators and then dozers and loaders. Many manufacturers offer machines operated by remote control. Extensive research aimed at developing robotic components and adapting the robots to practical service conditions is being conducted by various corporate laboratories throughout the world, including, e.g., the research and development units of the Caterpillar Corporation and Komatsu.

This research covers modeling of excavators as robotic manipulators [14.51], cognitive force control of excavators [14.52], soil mechanics with regard to ground-working tool interaction, kinematic and dynamic analysis of mechanisms, sensors enabling the determination of the position of mechanisms in a chosen location and in an absolute reference system, bucket trajectory optimization (leading to improved output and fuel economy) [14.47–50], and sensors detecting the presence of people within the machine's work range

and making it possible to determine the positions of the excavation's edge and the output removal vehicles.

An example of a remote-controlled machine is the pull shovel shown in Fig. 14.129. It is radio-controlled and has an operating range (distance from the control panel) of 1500 m. The working system's cylinders and travel drive are controlled by levers on the control console. Two video cameras, mounted outside the excavator, transmit a picture of the working area. A third camera inside the operator's cabin shows the indications of the gauges.

Fig. 14.129 Remote-controlled pull shovel

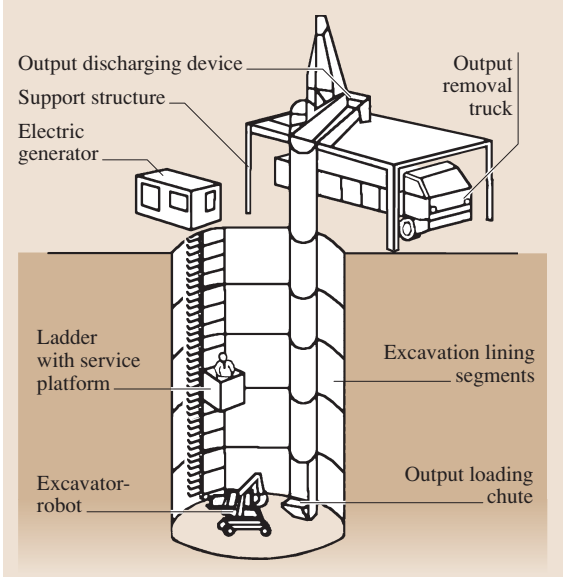


Fig. 14.130 Excavator with associated equipment for making deep point-excavations

Another example of a remote-controlled excavator is the excavator shown in Fig. 14.130 [14.53]. The machine is intended for making deep excavations, vertical transport of output, and loading it into transport means. A slewing body with excavating tools and containers or negative-pressure conduits for transporting output are mounted on the crawler chassis. The excavation's minimum diameter is 3 m.

The excavation rate and the output loading rate can be adjusted to the ground's properties. If soft rock is encountered, a milling drum can be mounted.

The excavator is used in mountainous terrain not easily accessible to pile-drivers to avoid the hazards (landslides, subsoil waters, toxic gases, and falling objects) to which the operator would otherwise be exposed.

Among point excavators there is a machine designed for work in hard rock [14.53]. Its rotary tool, mounted on a 0.7 m^3 undercarriage, is hydraulically driven. Boreholes 846 mm in diameter can be drilled.

Besides whole machines, computerized systems for controlling construction machines, allowing one to automate partially the operation of machines, are also available. Offered control systems can be installed on excavators, dozers, graders, and asphalt pavers. They incorporate tachometers, global position system (GPS) receivers or laser surveyor's levels. A simple example here is a control system for setting the position of the dozer's blade (Fig. 14.131) so that smooth, leveled soil surfaces can be obtained. A rotary laser on a tripod produces a horizontal reference plane. The control system with a signal receiver, mounted on the dozer, automat-

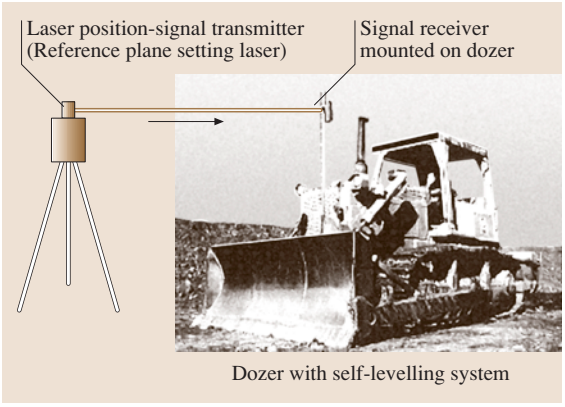


Fig. 14.131 Dozer with blade position controlling system for precision leveling of terrain

ically adjusts the elevation of the blade as the laser receiver follows the adopted reference plane.

A computerized control system for a grader works as follows. The task to be performed is stored in the memory of the machine's PC in accordance with a digital work execution scheme. The machine's actual location relative to the design data is determined by a tachometer or the GPS, which compares it with the design terrain elevation at a given point and on this basis sends signals to the machine's hydraulic system. For one such control systems, the accuracy of leveling is ±5–10 mm when a tachometer is employed and ±15–20 mm when a GPS system is used. Thanks to automatic control the number of passes of the machine, and so the work realization time and cost, can be reduced.

Besides control systems for single machines systems for the remote control of a set of earthmoving machines (Fig. 14.132) working in areas posing hazards to operators are also being developed [14.53]. Excavation and the loading and transport of the spoils are radio-controlled from a central control room located at a distance of 2–3 km from where the machinery is working. A vehicle with a radio relay station acts as an intermediary unit. The radio-controlled system incorporates the following subsystems:

- Transmission of stereoscopic images and graphics to a central control room.
- A bidirectional system for controlling vehicles and transmitting information about their position; the vehicles are located within a radius of 1 km from the radio relay station (Fig. 14.132).
- Audiovisual transmission providing information to the operators.
- Remote measurements using the GPS and an automatic tracking device (a three-dimensional laser positioning device).
- Monitoring the movement of the vehicles and the progress of the work and printing the results.
- Transmission of information about the excavation's cross-sectional dimensions and the positions and inclination of the vehicles.

Pneumatic Caisson Work

In pneumatic caisson work, because of the hazards (decompression sickness) to which the operators are exposed when moving from a pressurized space to atmospheric pressure, it is highly desirable to eliminate direct operation through automation [14.53]. Several methods of carrying out caisson earthwork have been developed in Japan [14.53]. One of these is a method of unmanned caisson work. The minimum diameter of a cylindrical caisson which can be used in this method is 8 m. If a rectangular caisson is employed, its minimum dimensions are 8.0×6.5 m. The preliminary work, which includes leveling of the ground to ensure even sinking of the caisson, is carried out using conventional earthmoving equipment. The system's common feature is remote control from ground level.

The equipment shown in Fig. 14.133 consists of a caisson scoop, a fast spoil-loading device, a caisson

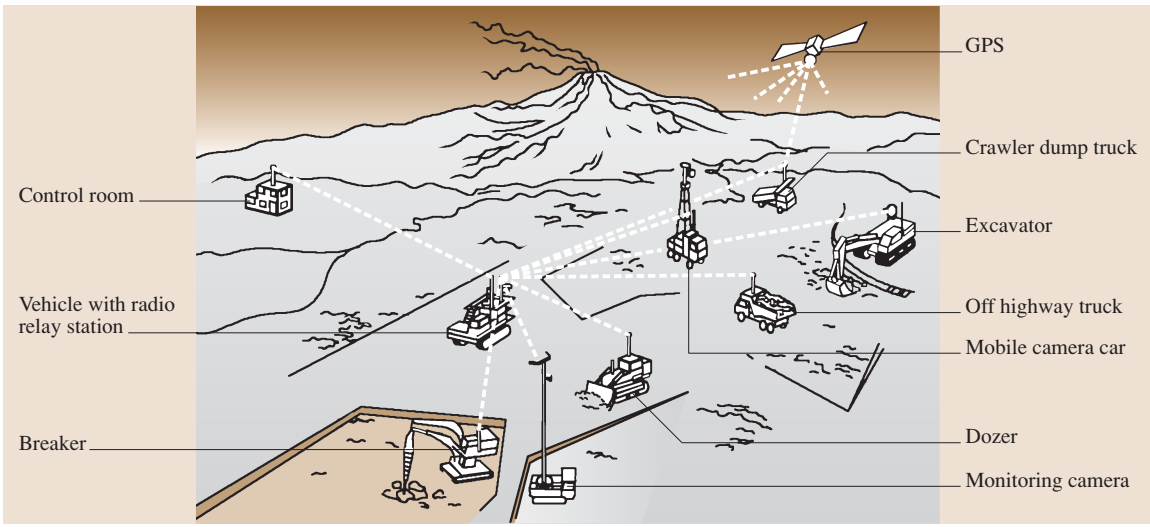


Fig. 14.132 System for remote control of earthmoving machinery

scoop control system, a measurement system, and a data transmission system.

The soil is excavated and transported by a specially designed caisson scoop that discharges its content onto a belt conveyor or the platform of a hoisting winch with a quick rotary-type spoil loader. Then the spoil is discharged into containers. The caisson scoop is radio-controlled and sensors installed on it monitor its operation.

Diaphragm Wall Construction

Diaphragm wall construction has been automated on construction sites in Japan. There are several commercially available systems for the erection of diaphragm walls. Automation covers excavation, the measurement of excavation execution accuracy, and quality control of the slurry stabilizing the excavation's walls and of the concreting process [14.53].

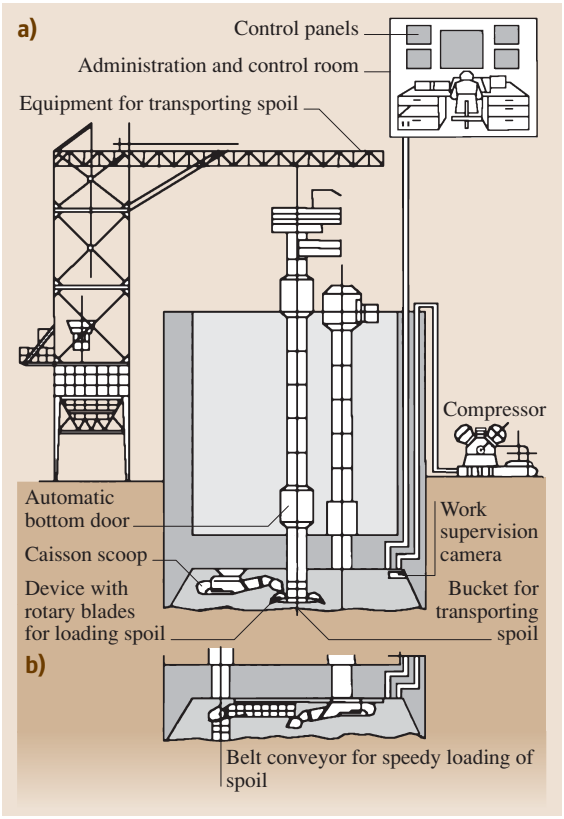


Fig. 14.133a,b Equipment used for unmanned caisson work: **(a)** equipment used in circular caissons **(b)** equipment used in rectangular caissons

An example of an automatic excavation system is the set of tools (Fig. 14.134) which was used for erecting a cavity wall of record dimensions (150 m deep and 2 m wide) in Izumiotsu near Osaka [14.53].

The system consists of a positioning system, an excavating system load control, and machine manual control by the operator on the basis of the information displayed on the monitor screen. The excavating unit, called an excavator, is suspended on a wheeled crane.

The measuring system consists of equipment located on two support structures and the excavators' instruments: adjustable guides, an inclinometer, a depth gauge, and a fuzzy controller to control excavation wall irregularity.

The position of the excavator is determined as a function of the horizontal shift of the two trace wires. The displacement of the trace wires is measured in an area of 100 mm^2 by a noncontact magnetic gauge with a junction measuring system.

The measuring system ensures accurate determination and correction of the excavator's position. For a 100 m deep excavation the excavating tool position accuracy is 30–50 mm.

Pile Driving

In the case of pile-driving machines, automation covers the control system ensuring the precise guidance of the working tool's end. A multijointed pile driver is shown in Fig. 14.135 [14.53].

The machine's main function is to drive in steel T-piles and sheet pile walls. In addition, various fittings, such as an earth drill or a vibratory hammer, can be attached to the multijointed work system's end. The machine is equipped with a computer-aided guidance system which guides the work system's end in the vertical plane as shown in Fig. 14.135. Before the new control system was introduced only a skilful operator simultaneously controlling the positions of the two arms had been able to make the work system's tip follow a linear trajectory. In the new system, which coordinates the movements of the two arms, control is effected by means of only one lever.

The computer-aided system of controlling the position of the work system's end ensures, by properly positioning the pile and maintaining its angle of inclination, that the pile is driven in accurately.

14.8.2 Automation of Concrete Work

Types of Concrete Works Covered by Automation

The automation of concrete works covers primarily the transport and distribution of concrete mix, and then the removal of the layer of corroded concrete from reinforced concrete structures, applying a new layer of concrete, the fabrication of reinforcement, the removal of irregularities in the surface of freshly set concrete, and the vertical shifting of formwork in the sliding erection process. Several applications of related techniques have been described [14.54]. The automation of concrete mix production in concrete batching and mixing plants is discussed in Sect. 14.3. Examples of the automation of the particular types of concrete construction tasks are provided below.

Transport of Concrete Mix

Most concrete mix transport automation solutions are found in the construction of dams [14.53]. The automation solution depends on the location and size of a dam. Figure 14.136 shows a diagram of an automatic concrete mix transport system used in the construction of a dam with a concreting work volume of 510 000 m^3.

The system covers the delivery of concrete mix from a concrete mixing plant to tipper trucks transporting it to the placing site. The fully automated system consists of the following devices: transfer car with a 4.5 m^3 capacity tank, two 4.5 m^3 capacity buckets, two cable winches with a hoisting capacity of 14.5 Mg, and two 9 m^3-capacity ground concrete hoppers. The transfer car draws concrete mix from the concrete mixing plant, transports it, and discharges it into one of the two buckets.

The operation of all the devices is controlled from a central control room. Information is displayed on a cathode ray tube (CRT) screen in the central control panel and, if necessary, the operator can intervene in the process. A basic requirement for the efficient functioning of the system is the precision positioning of the concrete buckets in the positions of concrete mix loading and unloading into hoppers. To ensure this, the bucket position is controlled in a spatial Cartesian coordinate system. The unloading of the buckets into hoppers is controlled wirelessly by means of systems with interlocking. The opening and closing of the hoppers gate valves is controlled by the truck drivers, who draw a specified quantity of concrete mix.

Besides computers and a program selector the other major components of the control system are gyroscopic sensors, optimeters, and coding units. Continuous speed control is employed in the cable winches' drives so that

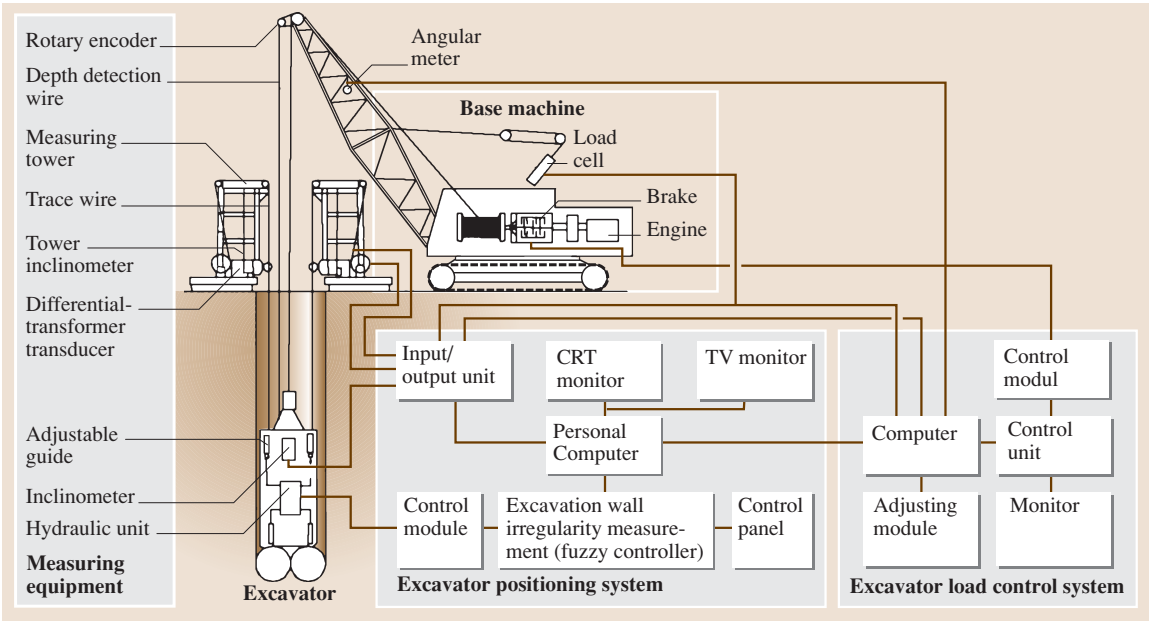


Fig. 14.134 Automatic diaphragm wall excavation system

the buckets can be quickly lifted and lowered, but positioned slowly.

Several concrete mix transport automated systems designed for the erection of dams are available on the Japanese construction equipment market.

Their common feature is the use of components such as transfer cars, cable winches, concrete buckets, hoppers, central control rooms, and radio-control systems. Instead of cable winches automated tower cranes can be used as the basic machines.

Distribution of Concrete Mix on Placing Site

Automation of concrete mix distribution is applied in the erection of large horizontal concrete structures such as floors and concrete bases. In this case, distributors fed by concrete pumps are usually employed [14.53]. Sometimes tower-mounted distributing booms are used, particularly in densely reinforced places where it is difficult to move distributors.

In order to be able to locate the delivery pipe's tip in the whole work area a typical distributor can perform the following work motions:

- Driving on a short (e.g., 3.2 m) portable rail-track
- Slewing of the whole delivery pipeline
- Mutual slewing of the particular pipeline segments
- Slewing of the flexible section of the delivery pipe

A concrete mix distributor with a capacity of 40 m^3/h (one of the smaller devices in this category) is shown in Fig. 14.137.

The distributor's delivery pipeline, which is 100 mm in diameter, consists of a steel pipe section and a hose, whose position can be changed in the vertical plane by means of the winch. The distributor can rotate at two points: at the junction with the travel section (on the track's axis) and at the junction with the rubber hose. In order to change its position the distributor is raised and turned by means of a jack lift and the rails are laid in a desired position. The distributor's whole delivery pipe can be rotated in the vertical plane in a range of ±35° and the hose in a range of ±108°. Moreover the elastic section can be raised up to 10° and lowered to 30°. The distributor is radio-controlled and one of the control modes is *partially automatic*, which eliminates the difficult simultaneous operation of two levers.

Among concrete mix distributors there is also a system for large concrete works with a four-jointed 20 m long delivery pipeline and an automatic concrete mix distribution system incorporating an automatically controlled tower crane [14.53].

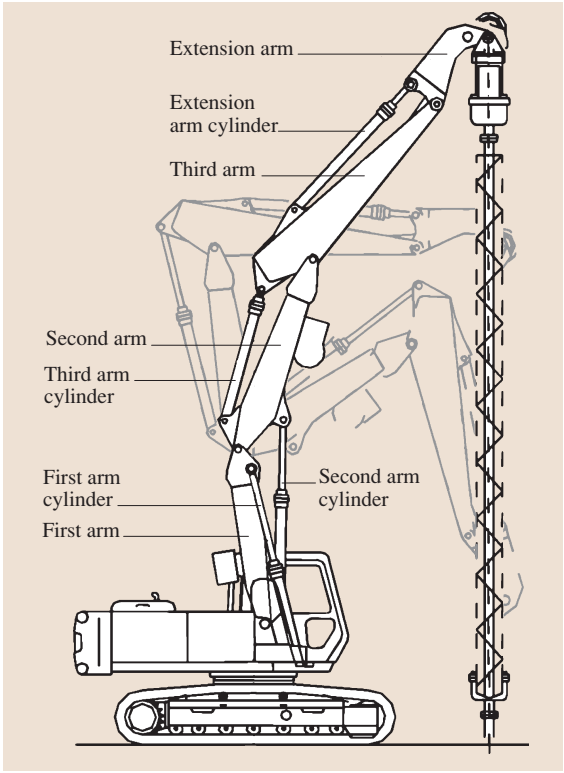


Fig. 14.135 Multijointed pile driver with automatic control of work system trajectory

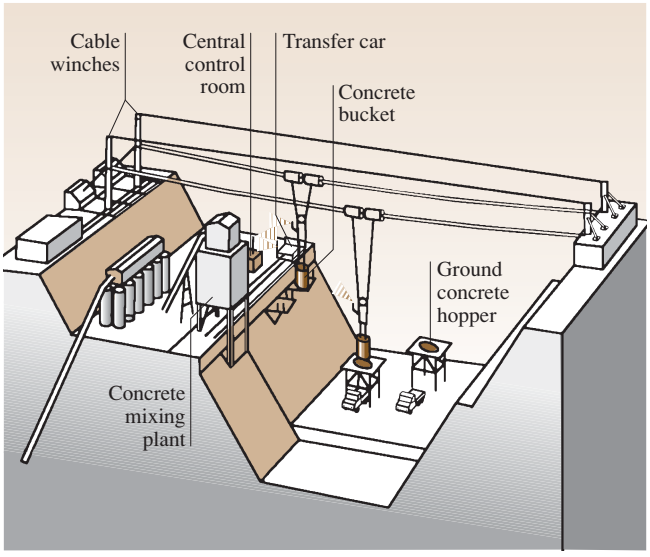


Fig. 14.136 Diagram of automatic system for transporting concrete mix on dam construction site

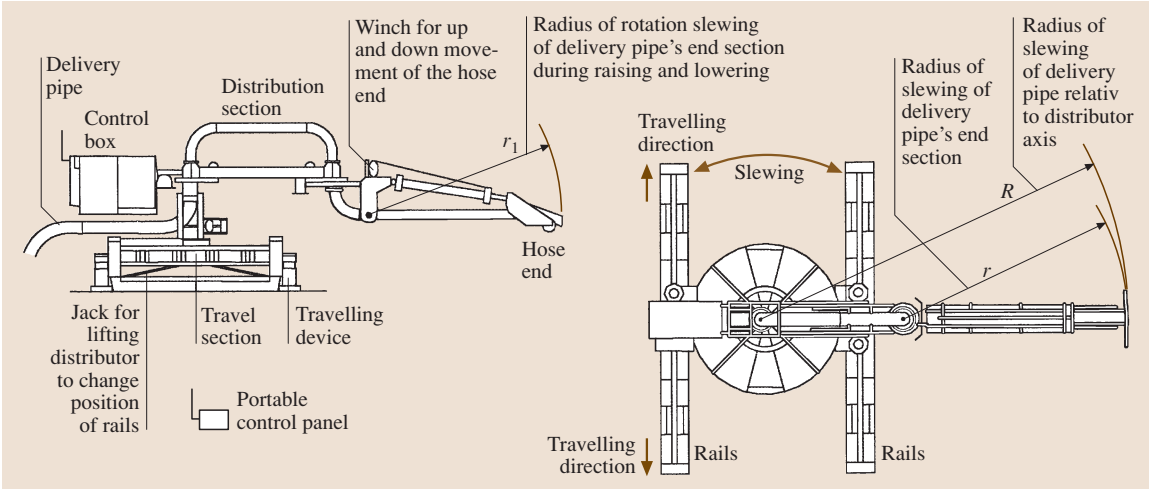


Fig. 14.137 Concrete mix distributor

For smaller structures, pumps with a distributing boom are employed [14.55]. Also here automation is introduced. One firm has developed *a computer controlled mobile concrete distributor* to distribute concrete mix on the placing site and better utilize the equipment.

Computer control makes it possible to position automatically the delivery pipe's elastic section from which concrete mix flows. In order to uniformly distribute concrete mix over the concreted floor one can select one of the three available modes of end section guidance. One of them is *follow-me*, shown in Fig. 14.138. The position of the pipe's elastic section is plotted in a coordinate system using a computer program. The operator sets the directions in which the elastic end is to move, directs the tip slightly in the desired direction (coordinates x, y), and toggles the switch up or down.

If the buyer wishes, the control system can incorporate the following modules:

- Fleet of vehicles management module (PD2000)
- Technical condition maintenance module (PARJS)
- Automatic stability control (ASC) module and automatic distributing boom control associated with concrete mix distribution

Concrete Spraying (Shotcreting)

The coating of reinforcement with a layer of concrete for anticorrosion protection purposes is used in new structures and for repairing old structures. The automation of this process is especially needed in the case of large surfaces, e.g., tunnels, pit shafts, and slopes in danger of sliding.

Research and development work aimed at automating shotcreting covers:

- Robots, incorporating devices for concrete mix proportioning, mixing, transport, and shotcreting; the robot shown in Fig. 14.139 [14.55] consists of a concrete mixing unit, a proportioning pump, and a remote-controlled robot arm. The robot arm is equipped with an automatically controlled oscillating nozzle which ensures accurate shotcrete application. Its action radius extends to 13 m, both

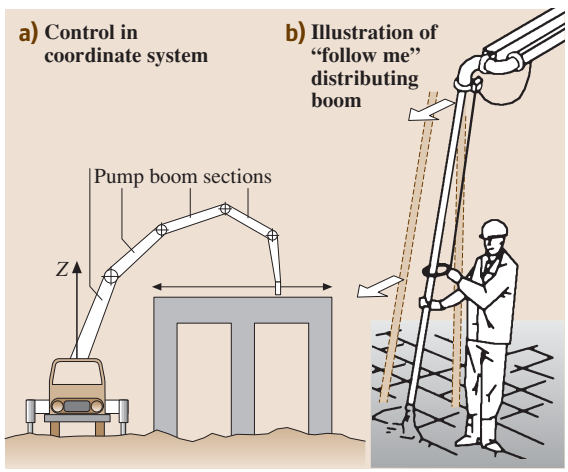


Fig. 14.138a,b Automatic control of delivery pipe's tip for concrete pump with *follow-me* distributing boom **(a)** control in coordinate system; **(b)** illustration of *follow-me* principle

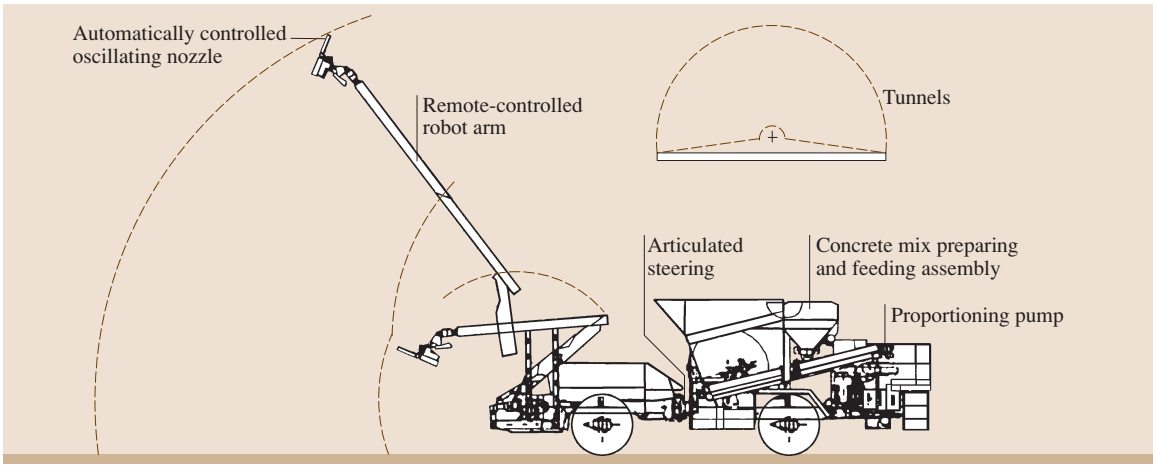


Fig. 14.139 Shotcreting robot

horizontally and vertically, and can easily be directed to cover all parts of the tunnel periphery. The concrete mixing unit consists of separate cement/aggregate bins with a total material volume of 5 m^3. The conveyance and mixing processes are effected by augers. All shotcreting is remote-controlled from a portable control panel, which enables the operator to choose the most effective position from which to work. The entire equipment is mounted on an articulated truck.

- A control system optimizes the amount of setting accelerant added to the concrete mix depending on the size of the concrete surface and the feed pump's delivery rate. The system can be used for any configuration consisting of a shotcreting device, a concrete feed pump, and accelerant feeder.
- Control equipment for secondary concrete lining placement. The equipment controls the quality of the shotcreted surfaces and automatically fills cavities at joints or cracks with concrete mix.

Removal of Surface Irregularities and Roughening of Freshly Set Concrete

The removal of surface irregularities and the roughening of freshly set concrete needs to be automated when erecting large concrete structures such as dams.

A robot for removing surface irregularities and roughening freshly set concrete is shown in Fig. 14.140 [14.53].

An automobile chassis with an attached assembly of four roughening brushes, a rolling sweeping brush, a threshold lift, and a waste bin is used as the suspension system.

The roughening brushes and the waste transport assembly can operate dry or with water supplied under pressure to the brushes and wastes removed by a vacuum pump [14.56].

Fabrication of Reinforcement

Reinforcement fabricating robots are used when erecting multistorey buildings with a reproducible system of storeys or on large construction sites such as nuclear power plants or underground reservoirs. A robot for the fabrication of reinforcement, mainly reinforcing bars, in both site steel yards and precast concrete plants is shown in Fig. 14.141. The objective is to reduce labor costs and ensure dimensional reproducibility of the fabricated elements. The robot automates simple, repetitive actions that steel fixers perform manually. Reinforcement with 4–6 m long longitudinal bars, 19–25 mm in diameter, can be fabricated. The robot consists of

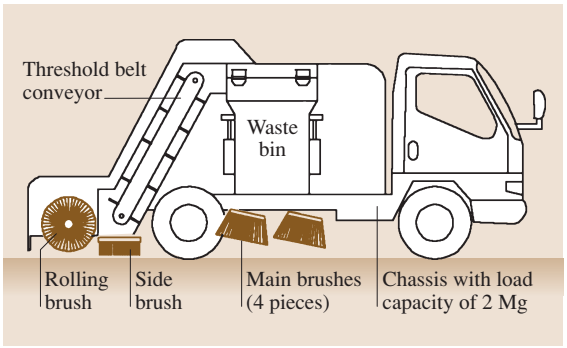


Fig. 14.140 Robot for removing surface irregularities and roughening freshly set concrete

the following assemblies: a base, an arm for feeding stirrups, an assembly for positioning stirrups, two (the upper and lower) automatic tying machines, a coil wire transfer car, arms supporting the (upper and lower) longitudinal bars, a power unit, and a control panel.

The main problem encountered when designing this robot was the precise longitudinal spacing of stirrups and securing them from shifting. This was solved by introducing a cam mechanism in which stirrups are positioned in the grooves of a plate that is shifted by the cam.

Other applications of robots to reinforcement fabrication include the following:

- Automated reinforcement pre-assembly line [14.53] designed for large structures and heavy rebars
- Automatic rebar bender and rebar column fabrication unit [14.53] designed for bending longitudinal rebars at an angle of 10–20° in six places and making a reinforcing cage

Removal of Damaged Concrete Layer in Reinforced Concrete Structures

Removal of a damaged concrete layer is used in repairs of structures such as bridges, pillars, tunnel walls, and dams, which have been damaged by the action of salts, environmental pollution and corrosion, or physical impacts. Two methods for the mechanized removal of a layer of concrete from a structure – a hydraulic method and a mechanical method – are distinguished.

In equipment operating on the hydraulic principle a highly pressurized (90–120 MPa) water jet penetrates the damaged layer of concrete and removes it at a rate corresponding to the operation of a few power hammers.

The advantages of the hydraulic method include: the possibility of adjusting the thickness of the removed layer through the rate of travel of the spray nozzle, eliminated dusting, a fixed magnitude of the acting forces (i.e., *healthy* concrete is not removed), and that rust can be removed from the reinforcement to be coated with a layer of concrete.

A limitation on the use of hydraulic equipment is ambient temperature, which cannot be lower than 0 °C.

Robots for removing a damaged layer of concrete by means of a high-energy water jet are shown in Fig. 14.142 [14.55].

The robot in Fig. 14.142a is equipped with a tower assembly kit providing an operating height 6 m. The robot shown in Fig. 14.142b removes damaged concrete from horizontal surfaces such as bridges and floors, but after reconfiguration can also be used on vertical surfaces.

The robot is controlled by a computer according to one of seven programs. All the machine travel parameters are set remotely from a control panel connected to the machine by a 4 or 6 m long cable.

In another robot operating on the hydraulic principle [14.53] a spraying nozzle whose work advance can be programmed through the association of slewing and advances was adopted. Such designs enable use in channels of circular and rectangle cross sections.

Accessories to hydraulic robots used for removing a layer of corroded concrete include 300 MPa feeders, distilled water tanks, and abrasive containers, in which cutting out holes in concrete structures becomes possible.

The number of robots available for the mechanical removal of a layer of corroded concrete is not large.

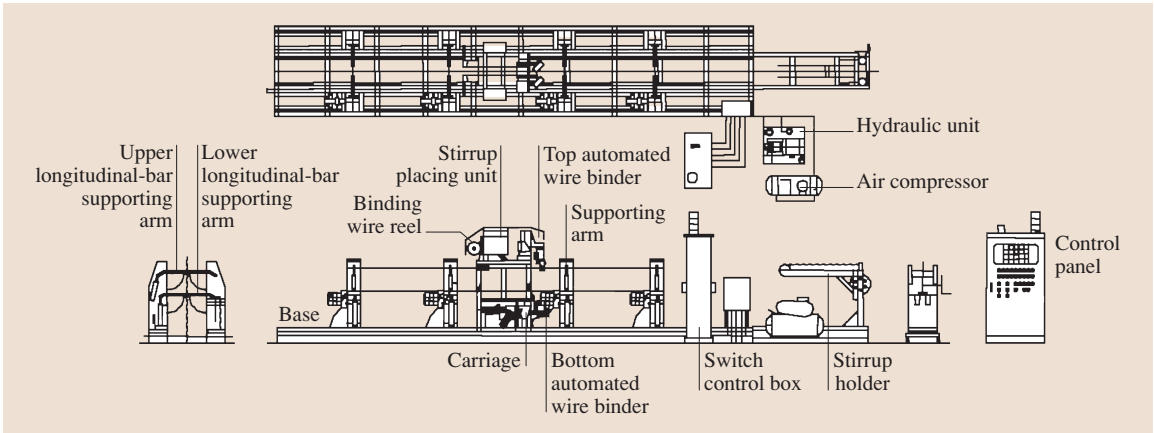


Fig. 14.141 Robot for fabricating rebars for reinforced concrete buildings

A *lining cutting robot* for work in tunnels resembling an ellipse sector in cross section is available [14.53]. The working tool is a rotary two-tool cutting head. The rotation of the head and the advance of the robot along the tunnel are controlled by a computer.

Forming of Concrete Structures for Sliding Forms

The automation of concrete structure forming mainly applies to sliding forms used in the construction of dams and structures such as chimneys, towers, silos, and bridge piers. A few methods of automating the erection of structures by sliding forms are described in [14.53].

14.8.3 Automation of Masonry Work

In European countries (Germany, the UK, and The Netherlands) research aimed at developing robots for masonry work has been conducted since 1991. This work has been focused on robots for erecting external and internal walls from aerated concrete and gypsum blocks.

A crane-manipulator (Fig. 14.143) for transporting and assembling 0.6×0.9×1.0 m aerated concrete blocks has been developed in The Netherlands.

The prototype robot shown in Fig. 14.144 was developed as part of the European Rocco project in 1995.

Research on the mechanization and automation of block-work wall erection focuses on the following equipment:

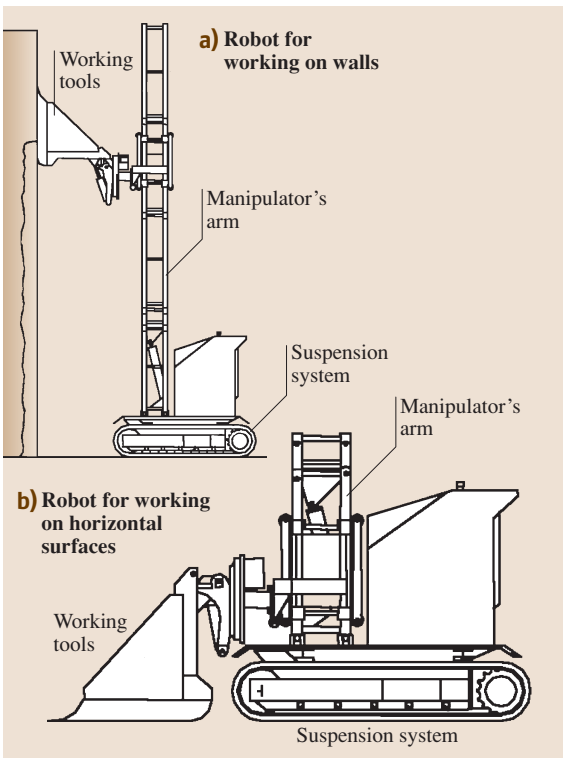


Fig. 14.142a,b Robots for removing damaged layer of concrete by means of a high-energy water jet: **(a)** robot for working on vertical walls, **(b)** robot for working on horizontal surfaces

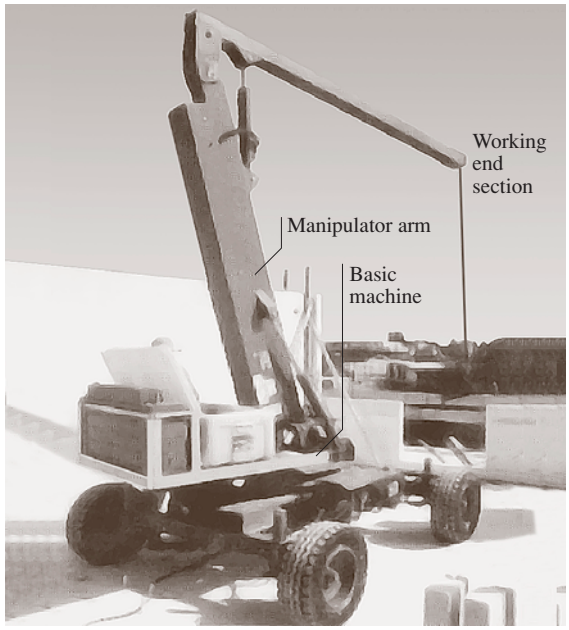


Fig. 14.143 Small-sized crane-manipulator for transporting and laying aerated concrete blocks

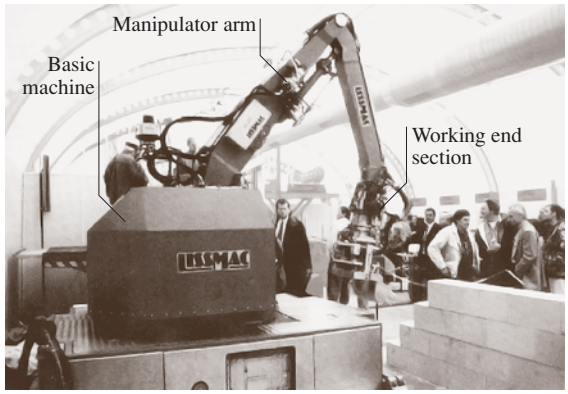


Fig. 14.144 Prototype robot for erecting walls of masonry, developed as part of the Rocco project

- Special bricklayer's platforms for adjusting the bricklayer's position to the height of the wall under construction
- Special winches or air-driven arms for compensating for the gravity of the blocks so that the latter can be manipulated weightlessly

Information on the numerous areas of research into robots for masonry work can be found in [14.45].

14.8.4 Automation of Cranes

The automation of cranes is connected with their specific use in a construction process. These are usually tower cranes (but also crawler hydraulic cranes) with an automatic control system, for transporting and distributing concrete mix and laying reinforcement.

A tower crane adapted for transporting and distributing concrete mix in housing construction is shown in Fig. 14.145.

The crane is a peculiar combination of the tower crane and the concrete pump's placing boom. The 32 m long four-sectioned boom can cover the entire work area. The final section of the boom concrete mix pipe is rigid (not flexible as in most cases) so that concrete can be placed accurately with no need to involve any workers.

Concrete mix is distributed from the level of the horizontally laid concrete mix pipe, whose position is remote-controlled. The elevation of the pipe above the concreted structure is constant and automatically maintained by means of computer control and a joystick. The position of the pipe's tip is specified in a Cartesian coordinate system on the basis of data (angles between the boom's particular sections received from sensors). In the case of joystick control, in order to maintain the boom horizontal, its position is automatically computed on the basis of the speed set by joystick and the current position of the boom. Joystick control allows one to move the first three boom sections in the vertical plane while the end section of the concrete mix delivery pipe remains horizontal. Depending on the application needs, the boom can be controlled automatically or manually.

Another example of the automation of cranes is an automated crane for transferring and arranging reinforcement (Fig. 14.146), intended for work on nuclear power plant sites. It can lay reinforcement at a rate of 0.05 m. No workers are needed to suspend the reinforcement bars, which can be arranged both horizontally and vertically by the system. The crane's specifications are:

- Maximum load: 1500 N
- Operating radius: 2.5–10 m
- Height of lift: 15 m

The crane is equipped with a special device for gripping rebars. The lifting and placing of the first rebar is manually controlled. This operation becomes automatic for subsequent rebars.

In [14.53] one can find more examples of the automation of crane control systems such as a crawler hydraulic crane for dam construction and a tower crane for the automatic distribution of concrete mix during the erection of a high-rise building.

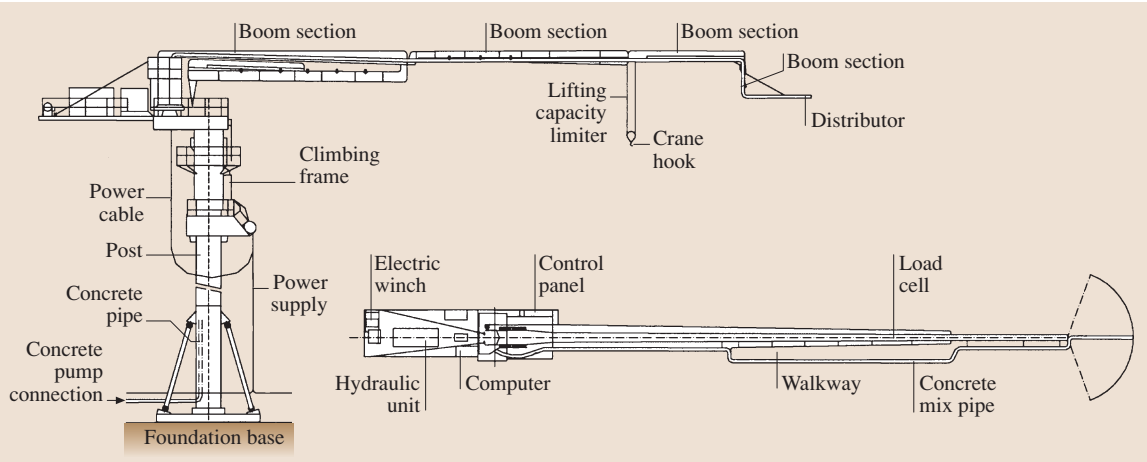


Fig. 14.145 Tower crane for placing concrete

14.8.5 Automation of Materials Handling and Elements Mounting by Mini-Cranes and Lightweight Manipulators

Mini-cranes and lightweight manipulators are used for materials handling, fitting building finishing elements, and transferring heavy construction equipment that cannot be moved manually. A radio-controlled mini-crane is shown in Fig. 14.147. The machine is intended for putting up aerated concrete block walls. This mini-crane with a three-sectioned telescopic boom has a lifting capacity of 500 N at an operating radius of 3 m. In its working mode the mini-crane is supported by four outriggers, whereas for relocation or transport it is mounted on a crawler chassis.

The crane can move up and down stairs and for transport can be fitted into a delivery truck. The crane's working dimensions are given in Fig. 14.147.

A lightweight manipulator designed for assembly work and transporting construction equipment inside buildings is shown in Fig. 14.148. Equipped with all kinds of attachments it can be used instead of interior work scaffoldings as a staging for ceiling work such as the fitting of lightweight beams, soffit, lightening, and heavy items. Its main advantage is that the position of the fitted element can be adjusted by means of the traversing and lifting gears operated from the handle attached to the manipulator's tip. Thanks to its small outer dimensions the manipulator can be moved through doorways and transported to different floors in the building's passenger lift. The manipulator has a lifting capacity of 150 daN and mass of 560 kg [14.53].

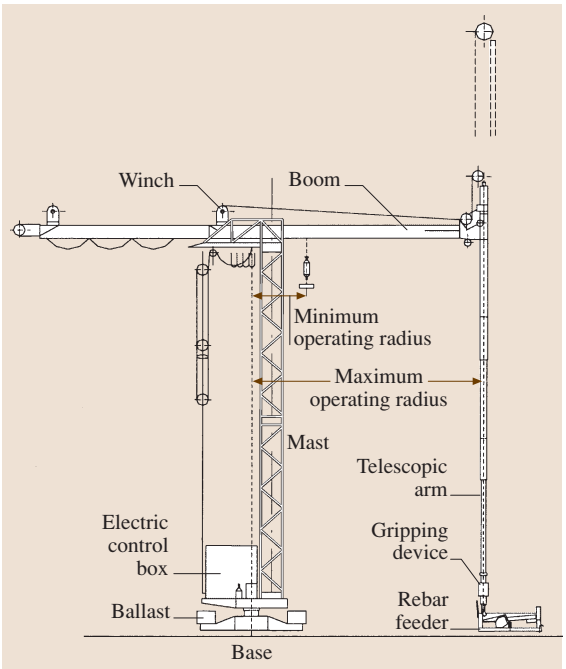


Fig. 14.146 Automated crane for transferring and arranging reinforcement

The offered mini-cranes and lightweight manipulators include:

- A mini-crane for fitting lightweight lining components, with a lifting capacity of 8000 N × 1.8 m, a lifting height of 5 m, a three-sectioned hydraulic boom, and a crawler undercarriage. It has an elec-

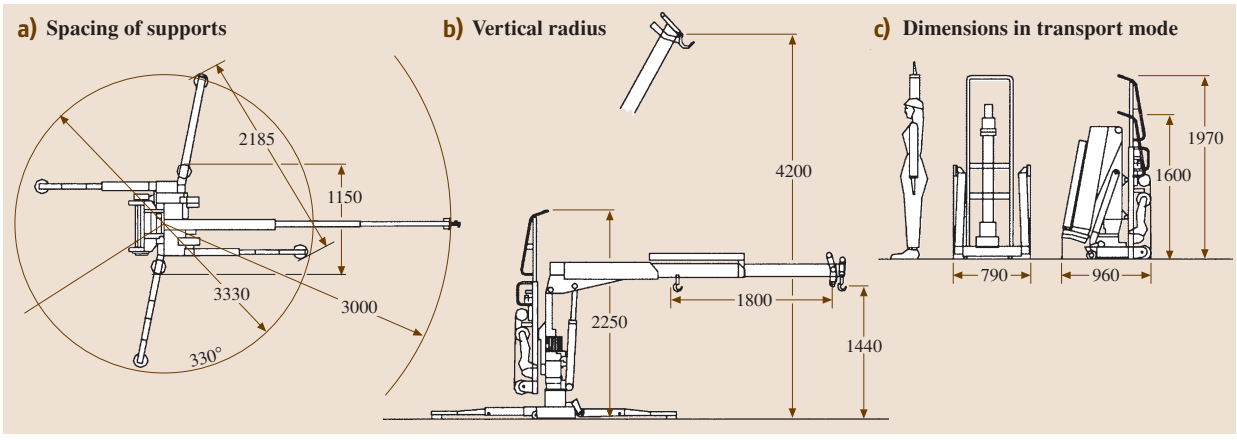


Fig. 14.147a–c Working dimensions of a mini-crane for erecting aerated concrete block walls: (a) spacing of supports; (b) vertical radius; (c) dimensions in transport mode

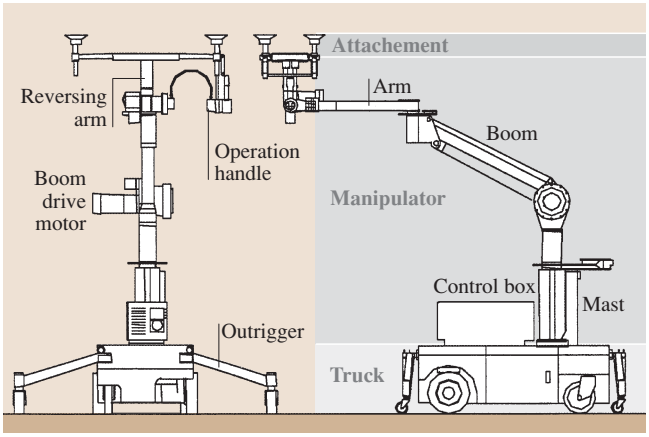


Fig. 14.148 Lightweight manipulator for finishing work

tric drive and a wired remote control. In order to make it possible for it to work in narrow rooms, the crane is equipped with a special mechanism, called the *dual-axis offset system*, which uses a combination of complex operations such as manoeuvring the winch while at the same time raising or lowering and extending or retracting the boom [14.53].

- A robot for transferring and positioning steel plates, weighing as much as 170 kg, used for reinforcing beams and columns in buildings located in seismic zones [14.53].
- A robot for fitting curtain walls made from panels, weighing as much as 250 kg, fitted between steel sections. The robot consists of a carriage with a wall panel gripping and transferring arm, positioning sensors, and a data-processing system which controls the fitting operation. The sensors on the arm determine the deviation between the appropriate position and the current one.

14.8.6 Automation of Construction Welding Work

Among construction welding work the welding of steel columns and beams is the focus of automation and robotization aimed at reducing welding work time and improving the working conditions of welders who, when hand welding on construction sites, are exposed to sparking, high temperature, and work at heights. The following devices for automating welding work on construction sites and prefabricating plants are available on the market:

- Robots for welding columns
- Robots for welding steel frames
- A device for automatic welding of girder braces

The robot shown in Fig. 14.149 is intended for welding columns together in steel construction. It executes horizontal multilayer welds. The whole column welding system consists of a robot with a wire terminal, rails attached to the column to allow the robot to travel, a control box, and a carriage for transporting the power source and a wire feeder.

The system also incorporates a welding control system. The shape of the weld between the columns is checked by laser sensors. The robot automatically selects the appropriate welding parameters from the system's database. By changing the shape of the guide rails and the control software the robot can be adapted to welding tubular columns.

Information about the steel frame welding robot and the device for automatic welding of girder braces can be found in [14.53].

14.8.7 Automation of Finishing Work

General Directions of Finishing Work Automation

A survey of the available literature indicates that automation and robotization are mainly applied to the following kinds of construction finishing work:

- Leveling, compacting, and smoothing concrete mix when making concrete bases

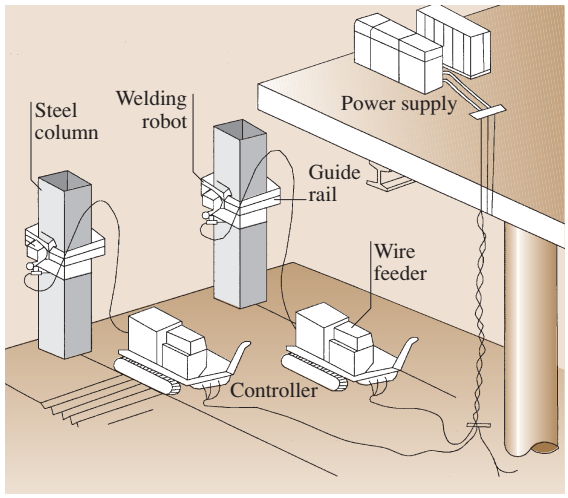


Fig. 14.149 Steel column welding robot

- Floating concrete bases and floors
- Painting
- Spraying fire-proofing materials on steel structures to protect them against damage by fire
- Laying tiles

Devices for automating the above finishing works are presented below.

Leveling, Compacting, and Smoothing Concrete Mix in Execution of Concrete Bases and Floors

Robots performing the function of a vibrating beam are used for precision leveling, compacting, and smoothing concrete mix when making concrete bases and floors.

The robot shown in Fig. 14.150 consists of the following units: a truss girder with a traveling mechanism, a working unit (composed of a leveling worm and a vibro-plate), saddles, and a control system. Leveling and smoothing are performed automatically by the working unit moving along the girder, i. e., perpendicularly to the movement of the robot. The maintenance of the base horizontal plane is controlled by a laser receiving set, an inclinometer, and a hydraulic cylinder, which automatically raises or lowers the working unit.

The robot is capable of leveling surfaces up to 15.5 m wide.

Another robot for screeding floors with a deviation from plane of ± 1 mm and a control system with two laser beam receiving sensors is described in [14.45].

Floating of Concrete Bases and Floors

The aim of floating is to impart smoothness to a concrete surface. Building floors on which fitted carpeting is to be laid must be floated.

Since the floating of concrete surfaces is often performed on construction sites, several types of robots controlled by radio or via an overhead conductor are offered [14.53]. Most of these are rotary floats with the working tools in the form of blades. In one exceptional model, six vibro-plates with a supersonic vibration frequency were adopted as the working tool. One of the models equipped with a traditional working tool in the form of rotating steel blades with an adjustable angle of inclination is shown in Fig. 14.151.

In the most common float models, the direction of the working travel is selected by appropriately inclining the floating disk, whereas the robot shown in Fig. 14.151 is equipped with a driving axle and an autonomous navigation system which allows it to determine its actual position and move along a programmed route (Fig. 14.152).

In order to program the robot it is enough to key in the dimensions a and b.

The robot is controlled via a wire. The travel control system consists of a microcomputer, gyrocompass, and a travel distance sensor. A touch sensor prevents the robot from bumping into obstacles. The robot's floating capacity is 500 m^2 per hour.

Painting

In painting work, robots are usually used to paint exterior walls. The next area where robots are employed is surface preparation for painting and surface preparation and painting combined.

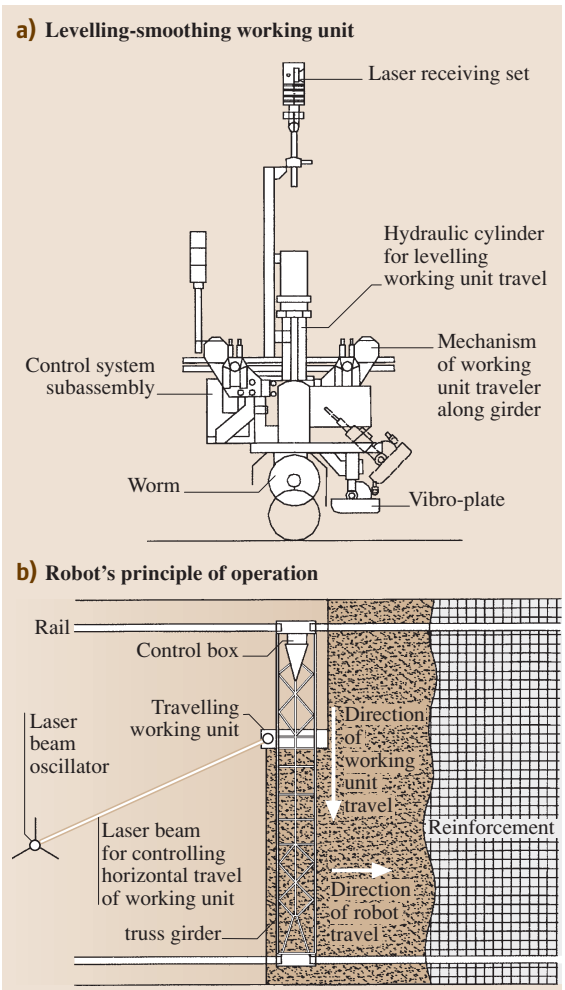


Fig. 14.150a,b Robot for finishing concrete bases and floors. **(a)** Leveling–smoothing working unit; **(b)** robot's principle of operation

One such robot for painting outer walls is shown in Fig. 14.153 [14.53].

The robot is mounted on a hanging scaffold and moves up or down and crosswise together with the scaffold.

Besides the scaffold and its drives and the robot itself, the system includes a control box, a control panel with buttons, monitoring equipment, and a paint-feeding unit.

The robot maintains a constant distance between the spraying nozzle and the building's wall. In order to apply paint uniformly, the nozzle's oscillatory motion is also controlled. To increase its painting capacity the robot is equipped with two spraying nozzles. The robot's average capacity is 200 m^2/h.

External-wall painting robots differ in their adaptation to construction site conditions (high-rise buildings with a complicated shape, or low buildings with simple design), and in their methods for reducing paint mist dispersion or eliminating unpainted areas. Paint can be applied by spraying nozzles or a roller. The preparation of a surface for painting consists of cleaning by means of a rotating cloth disk, with the dust being sucked off into a bag filter.

In the case of some robots, a surface is cleaned for painting by means of wire brushes or a shot-blast machine [14.53].

A robot for removing old paint prior to renovation painting removes paint using a jet of water flowing out under a pressure of 150 MPa. A pressure generator and a spraying nozzle are placed on a hanging scaffold attached to the wall by means of suction cups [14.53].

Spraying of Fire-Resisting Materials on Steel Structures

For fire protection the building's load-bearing structure must be coated with a layer of an insulating material – usually a 25 mm thick layer of rock wool. Because of the harmful effects of rock wool dust on the human body, workers manually spraying the semiliquid insulating mass must wear masks to protect them from inhaling harmful dust and wear special protective clothing.

Therefore it is vital to automate the insulating mass spraying operation and eliminate any direct participation of people in the application of the protective layer.

Two robots for coating structural beams with fireproofing mass are shown in Figs. 14.154 and 14.155.

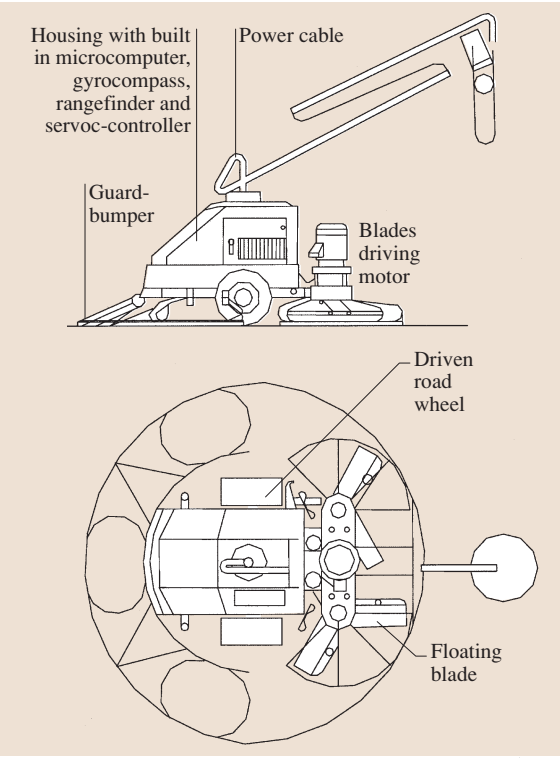


Fig. 14.151 Robot for floating concrete slabs

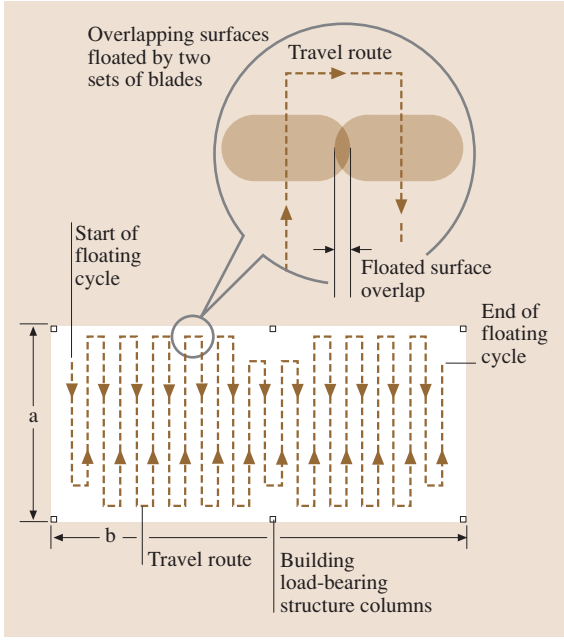


Fig. 14.152 Route of programmed robot travel in rooms with dimensions $a \times b$

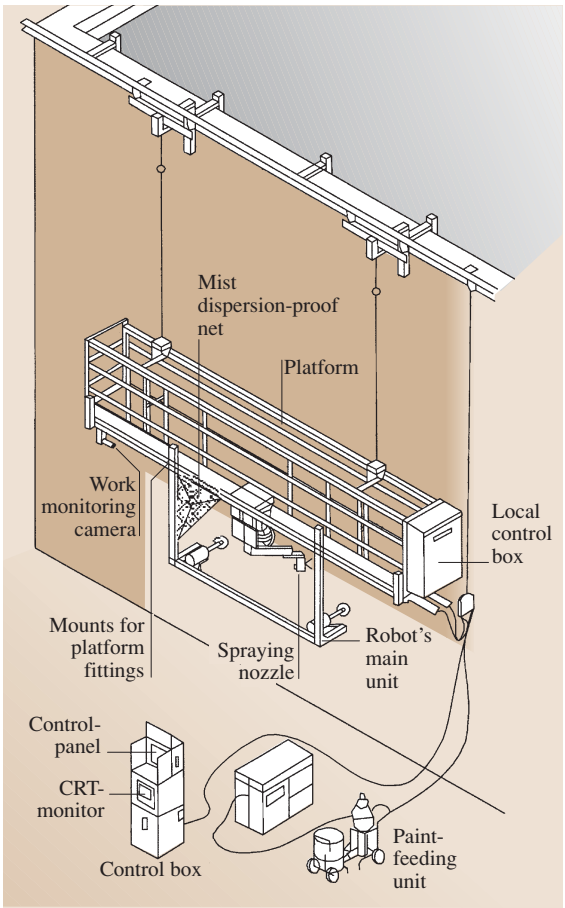


Fig. 14.153 Robot for painting exterior walls

The robot in Fig. 14.154 consists of two main fire-resistant mass preparation units (Fig. 14.154a) and the fire-resistance mass spraying robot proper (Fig. 14.154b). Once the beam's elevation and length are keyed in the robot starts moving, sensing its position by means of an ultrasonic sensor and spraying fireproofing mass on the section's bottom and sides. The spraying nozzle can be raised to a height of 2.8–4.4 m by means of a screw elevator. Special software makes the entering of robot operation data simple and easy.

The robot shown in Fig. 14.155 is intended for work in narrow rooms that are inaccessible to equipment operators. It consists of a carriage, an articulated arm with a spraying nozzle, and a control unit. The carriage has two sets of wheels: one for moving the robot longitudinally and the other for moving it transversely relative to the beam being sprayed with fireproofing mass. The articulated arm can rotate in both the horizontal and vertical plane and it can be moved to a distance of up 1500 m along the carriage's platform. Computer control enables the automatic combination of working motions for guiding the spraying tip.

Besides the described robots for spraying fireproofing mass onto structural beams one should also mention robots based on a typical industrial robot. An articulated arm with a spraying tip makes it possible to spray fireproofing mass on both horizontal and vertical surfaces [14.56].

Laying Wall Tiles

Laying tiles to form wall linings is an operation that is difficult to automate, mainly because of the requirement

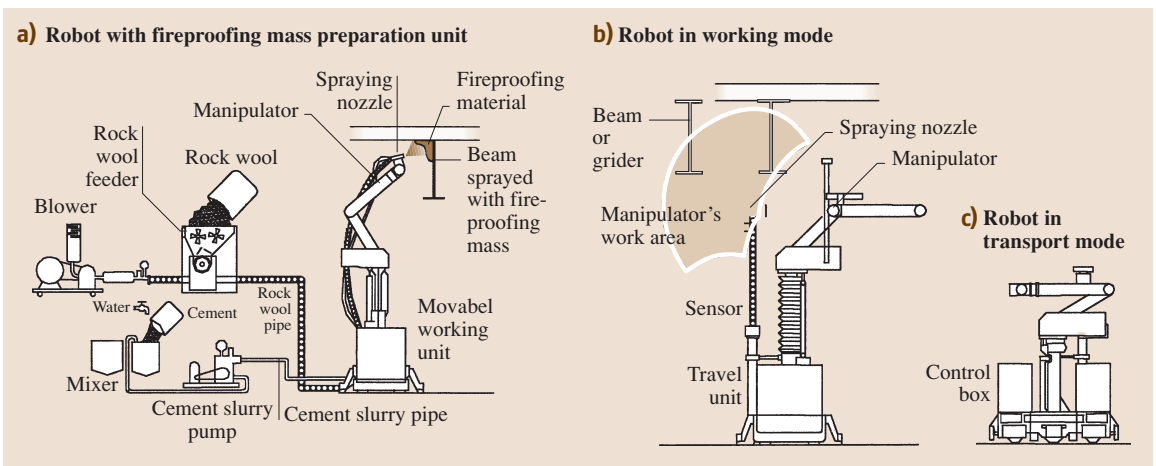


Fig. 14.154a–c Robot for spraying fireproofing material: (a) robot with fireproofing material preparation unit; (b) robot in working mode, (c) robot in transport mode

to spread adhesive or cement, which requires high positioning precision. Nevertheless, because of the large share of lining work in finishing work, attempts are being made to automate this field. Progress has been achieved by employing a rubber belt conveyor with suckers so that several tiles can be laid simultaneously. Tile-laying robots are mainly used for covering with large surfaces of exterior walls with tiles. A tile-laying robot moves on rails fixed to scaffolding. A robot for laying tiles on the building's facade, developed jointly by several Japanese companies, is shown in Fig. 14.156.

The robot is intended for laying 227×60 mm (8–15 mm thick) tiles on traditional mortar. Its daily capacity is 14 m^2. For comparison, a craftsman is able to lay 7 m^2 in this time.

More information about the robot can be found in [14.53].

14.8.8 Automated Building Construction Systems for High- and Medium-Rise Buildings

Automated building construction systems (ABCS) for high- and medium-rise buildings were developed in Japan as a measure to alleviate the labor shortage in the construction industry. Fourth-generation robots, designed according to the principle of design for robotic construction (DfRC) integrating robot design, building erection, and the materials, are employed for the realization of buildings in such systems.

Prefabrication of structural elements and adherence to the schedule of materials delivery to the construction site are of vital importance.

The benefits from using ABCS include: labor saving, improvement in work safety, and a reduction in construction time, stemming from the minimization of labor and protection against inclement weather (the working platforms are provided with roofing).

The following features of ABCS can be distinguished:

- *The automated transport of building structural components*, i. e., columns, beams, flooring slabs, and exterior walls from a storage place to the construction site (where they are built)
- *The use of prefabricated components* so that forming and working on the construction site are eliminated
- *Just-in-time delivery* of the required components
- *Automatic positioning and fixing of components*: the prefabricated components of columns and structural beams, floor panels, outer walls, and modular piping are automatically positioned in the appropriate

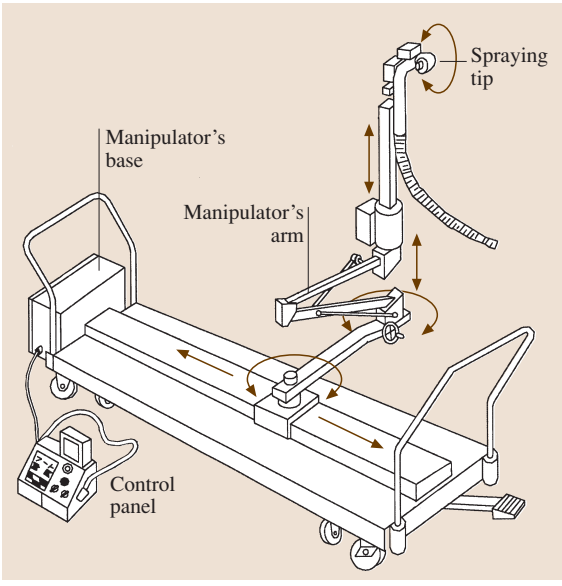


Fig. 14.155 Robot for spraying fireproofing rock wool mass in small rooms

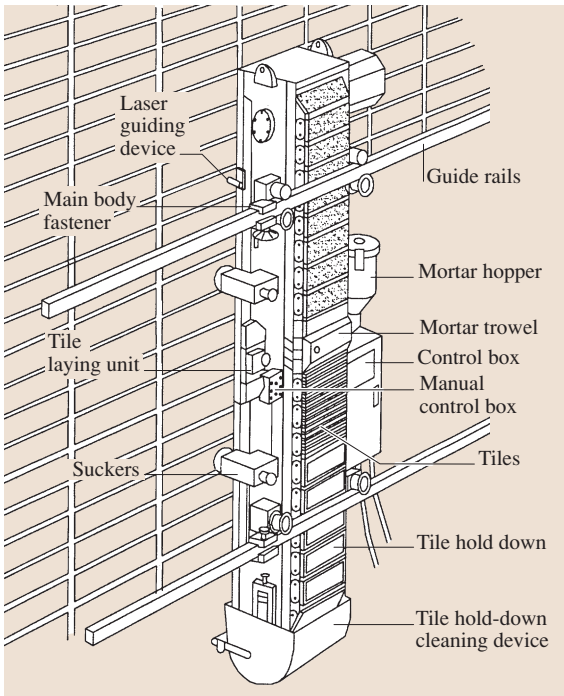


Fig. 14.156 Robot for laying tiles on building elevation

places, welding robots weld the structural components together, and the exterior walls are fixed by quick-connect fittings
- *Raising (by means of hydraulic lifts) of the support structure* with the suspended overhead cranes (the working platform) for the construction of the next storey
- *A building design that takes into account work automation*, sot that all building components and their joints lend themselves to automated transport and assembly

The existing ABCS can be divided into the following three groups:

- Those employing a special support structure (situated above the building structure under construction) for suspending overhead cranes and hoisting winches [14.57]
- Those using the roof structure or the top floor as the working platform for fixing transport-assembly equipment
- Those consisting of the construction of successive storeys on the zeroth level and the raising of the constructed part of the building

In all the above construction systems it takes 5–8 days to build one storey.

ABCS are used for erecting mainly steel-frame buildings, but also prefabricated reinforced concrete buildings [14.58].

Only a few of the many ABCS have stood the test of time and proved economically and technologically viable.

Selected ABCS are described below.

ABCS Employing Special Support Structure for Suspending Overhead Cranes and Winches

A construction system based on a special support structure for suspending overhead cranes and traveling winches during the erection of a steel-frame building is shown in Fig. 14.157.

The basic equipment for work realization in this system includes the support structure with a suspended

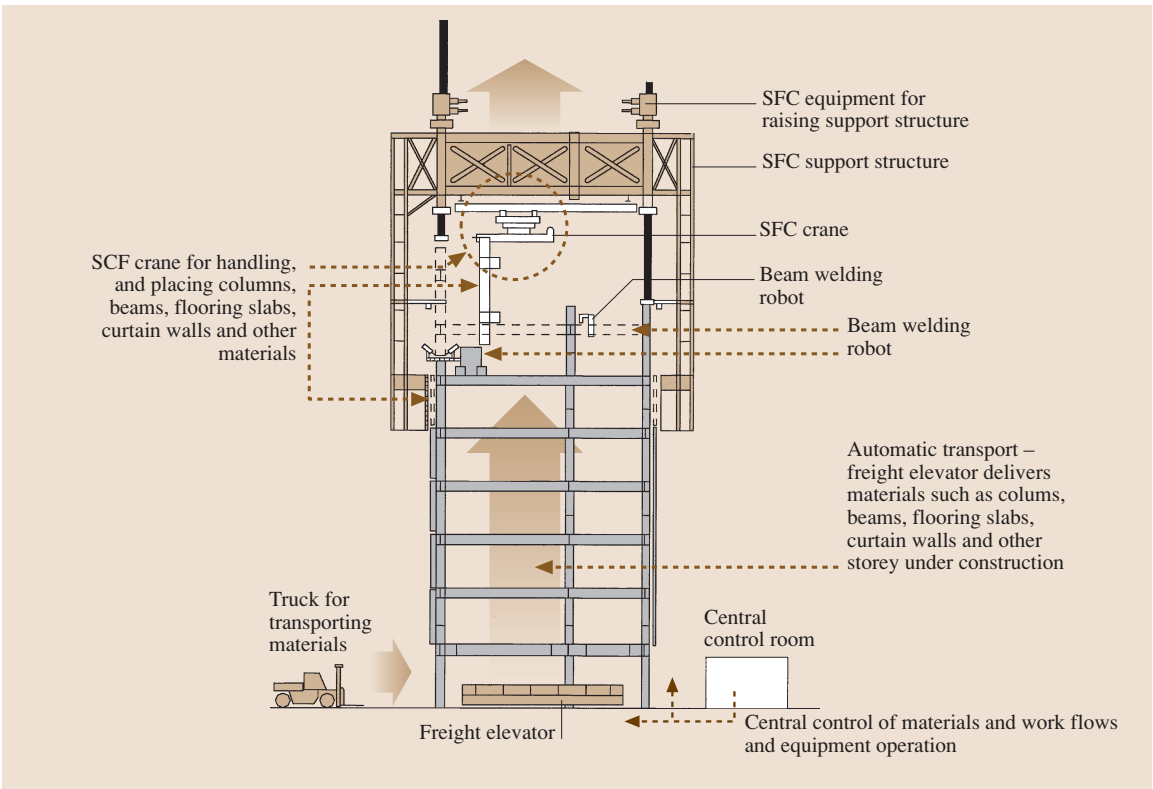


Fig. 14.157 Schematic of ABCS for erecting steel-frame buildings. SCF – the support structure with a set of equipment located on the top storey is called a super construction factory (SCF)

overhead crane and robots for welding the building's columns and structural beams. The support structure's weight is carried by the building under construction. The crane's operation is automated and coordinated with an elevator delivering materials in the appropriate sequence. Welding robots are attached to a floor or beams by one-touch fasteners. All operations are controlled from a central control room.

An ABCS for constructing buildings from prefabricated reinforced concrete slabs, referred to as *Big Canopy*, is shown in Fig. 14.158. This system has been developed by the Obayashi Corporation of Japan.

The special support structure, called a canopy, is supported on four corner posts.

Three assembly cranes and winches, each with a hoisting capacity of 7.5 Mg, are suspended from the canopy. Structural members are transferred to the storey under construction by a traveling crane, a fast building elevator, and a combination of overhead cranes and winches in order to provide access to the entire assembly area. The canopy is raised every two storeys.

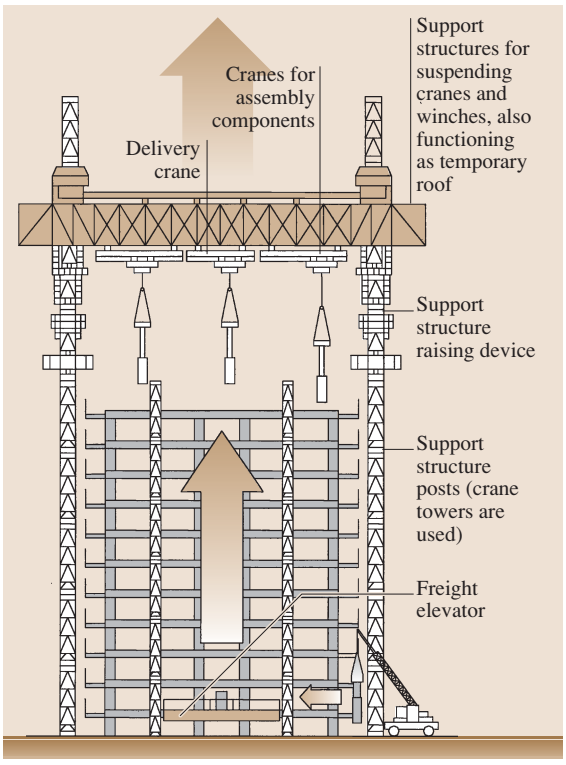


Fig. 14.158 Schematic illustrating the erection of a reinforced concrete building using the Big Canopy system from Obayashi Corporation

Finishing materials are transferred to the places where they are to be built-in during the erection of the load-bearing walls. The main part of the canopy is dismantled once the building reaches its full height, while its perimeter frames are jacked down and dismantled on the ground.

The automatic control covers the delivery of building components from the stacking yard, their transfer within the building, and their assembly. The components are identified by plates bearing a bar code which includes the position in the building and basic specifications.

ABCS Using the Roof or Top-Floor Structure as a Working Platform

Construction systems using the roof or top-storey structure as the working platform constitute the most numerous group of ABCS. A schematic of one such system which uses the top floor as the platform for the assembly of the steel load-bearing frame and reinforced concrete flooring slabs is shown in Fig. 14.159. The system is called the Shimizu manufacturing system by advanced robotics technology (SMART) [14.53, 59].

Construction starts with the assembly of a roof (hat-truss) and a special support structure for suspending overhead cranes and traveling winches The two structures are assembled on the ground and then jacked up using a system consisting of four tubular pillars equipped with lifting gears. A single lifting gear consists of two rings (the top one and the bottom one) and three hydraulic cylinders between them, each with a lifting capacity of 120 Mg and a working pressure of 12 MPa.

Once the roof is raised, the first storey is built. When it is completed, the roof with the support structure is jacked up one storey. Then the lifting system is also shifted up and supported on the load-bearing beams of the previously erected storey. As the individual storeys are constructed, structural building components (reinforced-concrete flooring slabs, interior and exterior walls, and utility systems) are fitted. Even though one storey weighs about 1200 Mg, it takes only 1.5 h to jack up the roof with the support structure and shift up the lifting gear.

In order to provide protection against bad weather the roof and the periphery of the storey under construction are lined with screens.

SMART embraces the automatic handling and assembly of building components, welding by robots, the raising of the roof and the structure for suspending overhead cranes and winches, protection against

adverse weather conditions, and a computer building construction control system. Because of its integration of the automation of the particular kinds of construction work with control systems, SMART is considered to be a state-of-the-art system.

Besides the described systems, several other systems based, among others, on the idea of constructing consecutive storeys on the zeroth level and successively lifting up the constructed parts of the building have been developed. Besides Shimizu and Obayashi, a number of other Japanese engineering construction firms have developed their own proprietary systems with similar functionality, e.g., Taisei, Kajima, Takenaka, Fujita, and others. Buildings up to 15 storeys high are erected with the use of these systems.

14.8.9 Automation and Robotics in Road Work, Tunneling, Demolition Work, Assessing the Technical Condition of Buildings, and Service-Maintenance Activities

This section provides a general overview of systems developed in this area and the reader is referred to the additional technical literature on the subject for further details.

Automation of Road Work

Automation and robotization of road construction and maintenance presents unique sets of challenges and opportunities, as described in [14.60]. Roadwork includes earthwork, concrete work, and other works connected with road trench cutting, road-base making, and the placement of pavement. In this section, only the machinery connected with the automation of pavement work is presented. It seems that automation in this field has focused on laying of asphalt mixture and road profiling.

Robots with all the pavement work operations automated are available in industry [14.53, 55]. The operations include transfer of asphalt mixture from delivery vehicles, feeding and spreading of asphalt, steering of the machine along a fixed route, paving rate control, and the start–stop control of all the operations. The asphalt paver operator does not need to pay attention to the loading of fresh asphalt mixture or driving the machine along a fixed route; their main task is to watch the screed to ensure that a high-quality pavement is obtained.

The asphalt paver's computer control system controls the following parameters:

- The thickness of the placed layer of asphalt mixture – the current paving thickness is displayed on the cab's monitor.
- The road profile – the operator can automatically change the paving thickness at one end or the other end of the screed.
- The uniform feeding of fresh asphalt mixture to the screed.

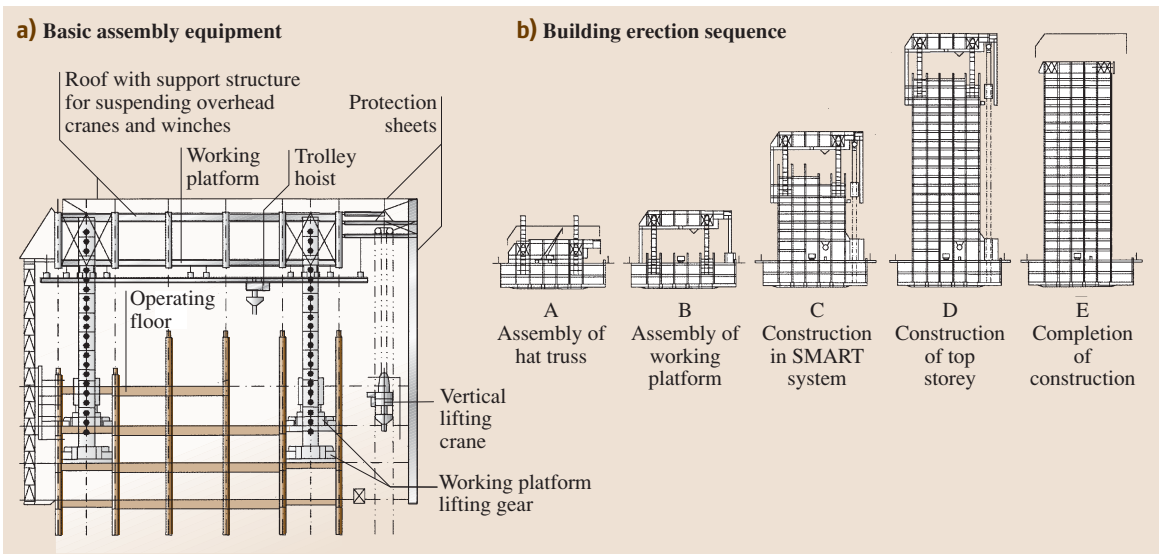


Fig. 14.159a,b SMART system for erecting steel-frame buildings: **(a)** basic assembly equipment; **(b)** building erection sequence

- The amount of crown.
- The uniform distribution of temperature on the screed.

Another machine is a hot in-place recycling machine intended for laying reclaimed asphalt pavements (RAP), which:

- Mills the pavement undergoing renovation, heated up by a preceding machine
- Prepares new asphalt mixture, which includes: reclaimed asphalt, fresh asphalt mixture delivered from an asphalt mixing plant, and chemical additives
- Lays a new layer of asphalt mixture

The hot in-place recycling machine consists of two separate, self-propelled units loosely connected by a conveyor. The set's front unit is composed of a feeder of fresh asphalt mixture delivered from an asphalt plant, a milling subunit, and a subunit for spraying additives. The rear unit consists of a reclaimed-asphalt pavement feeding and weighing subunit, a fresh-asphalt mixture batching unit, a mixer, and a subunit for laying new pavement. The two units move independently towards the destination, but operate in tandem.

The machine's automation covers: distance maintenance, the harmonization of the two units' driving speeds, the preparation of asphalt mixture, and feeding the mixture to the screed and laying it.

An example of a high level of automation is a road milling machine equipped with an automatic cutter control system (ACCS). With ACCS the machine can work in two modes:

- Contour-following mode, in which a constant cutting depth relative to the road surface is maintained
- Longitudinal and transverse contouring mode, in which the cutting depth is progressively adjusted to the required lateral differential

The cutting depth can be accurately adjusted to compensate for longitudinal and transverse unevenness in the pavement.

The cutting depth data for the road section to be milled are entered into the onboard computer, which automatically controls the position of the working tool in the selected operating mode.

Automation of Tunneling Work

The automation of tunneling is essential because of the hazard to people and the difficult working conditions, which are similar to those in underground mining.

The main aim of automation is to eliminate the presence of workers in the danger zone where they could be exposed to headwall landslides and inrushes of underground water in a confined space. Tunnel construction owes its rapid development to automation. The construction of municipal transport tunnels in urban areas, mini-tunnels for utilities (particularly for sewage pipes), and mountain tunnels for intercity transport is a major factor in the economic development of cities and regions. The general trends in the automation of tunneling work are presented below. Extensive information on tunneling machines and equipment can be found in [14.53].

The automation of shield tunneling covers the following operations:

- The automatic transport and assembly of prefabricated tunnel lining units, for which several systems of automatic lining transport and assembly have been developed for the different kinds of lining joints used
- The control of shield advance along a programmed route
- The complex automation of: tunneling, tunnel heading stability and shield advance control, output transport, lining assembly, and filling the space behind the lining with cement
- Shotcreting the lining's top layer reinforcement.
- Fabrication of reinforcement for the monolithic tunnel lining layer.
- Screeding of the top layer of the tunnel's invert.
- Transport and assembly of prefabricated concrete slabs for the tunnel's railway subgrade.
- Diagnosis of tunneling shield defects.

Most automation solutions are found in the areas of automatic transport and assembly of tunnel lining components and control of shield advance along a fixed route.

The automation in the construction of mountain tunnels covers the following aspects:

- Tunnel boring – when the start button is pushed the boring machine bores the ground in the tunnel's face, maintaining the tunnel's design cross section and following the fixed route with a high degree of accuracy. The available machine designs are capable of boring tunnels in hard and semi-hard rock without using explosives.

- The construction of a pre-lining (the pre-lining support method – PLS) to protect the tunnel face and adjacent structures against ground settlement; the pre-lining is made by excavating the soil around the tunnel face circumference and filling the slit with concrete; the PLS machine is equipped with a set of five augers or a chain cutter.
- The construction of another monolithic tunnel lining, using the vibration method to compact concrete mix.
- Tunnel face determination and marking prior to tunneling.
- Shotcreting of the lining's surface layer
- Determination of the position of explosive charges in the tunnel face
- Measurement and investigation of the fractured region and the strata in the ground ahead of the tunnel face
- Ventilation of the tunnel during its construction

A discussion of tunneling work automation should include the automation of mini-tunnelling. Mini-tunnelling is used in the construction of municipal sewerage and water-supply systems, gas grids, and so on without open excavations and the associated traffic problems. The technique consists of sinking two shafts (working chambers) at the start and end of the pipeline's route, boring a mini-tunnel by means of a tunneling shield, removing the excavated material, and forcing through pipes to form the mini-tunnel's lining (Fig. 14.160). The spoils can be removed mechanically (on conveyors) or hydraulically, being transported in suspension (partially recoverable). Equipment for boring mini-tunnels 250–4000 mm in diameter, running straight or curvilinearly, is commercially available. Mini-tunnels longer than 1000 m require intermediate working chambers. This technique can be used even in rocky soils.

Automation of Demolition Work and in Repairs of Building Structures

The aim of the automation of demolition work and in repairs of building structures is to increase productivity, reduce costs, and improve the working conditions of the laborers by improving their safety and reducing their physical effort. The devices presented below are used for crushing bricks, stones, and concrete, repairing cement kilns and metallurgical furnaces, and removing the surface layer of concrete in structures undergoing renovation. Several robots have been developed for these purposes. Depending on the size of the reduction tools, they can be divided into two groups:

- Robots equipped with mechanical tools, e.g., a hydraulic hammer
- Robots equipped with hydraulic tools using a high-pressure water jet

The two groups of machines have different ranges of application. Robots equipped with a hydraulic hammer and other fittings are designed for crushing building materials and for repairs, whereas those using a wa-

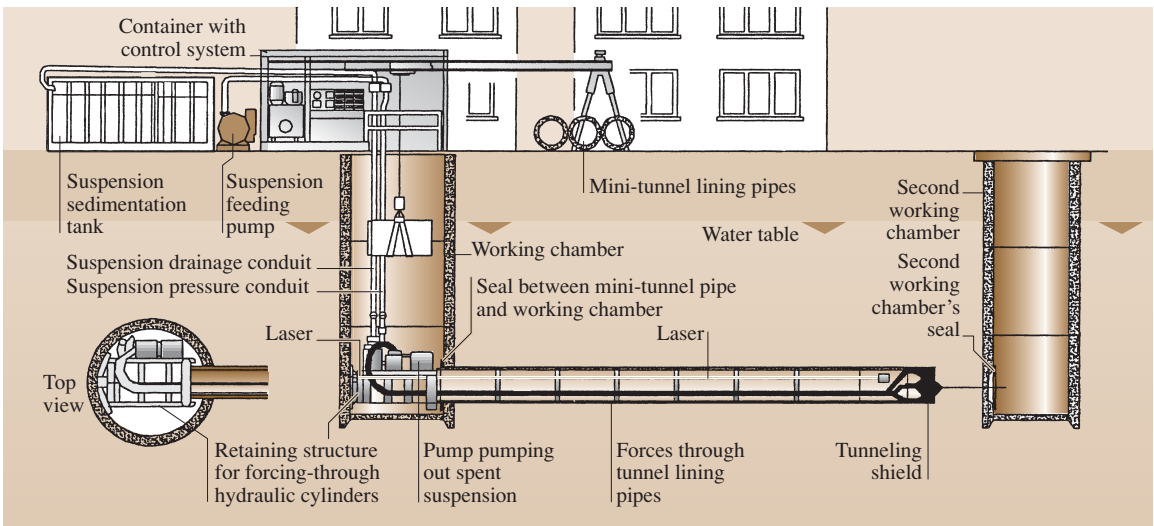


Fig. 14.160 Illustration of a mini-tunneling technique:

ter jet are intended for removing a damaged layer of concrete.

In the group of robots equipped with a hydraulic hammer one can distinguish, depending on the installed power and the tool's reach, small, medium, and large machines. Machines for small jobs are equipped with a 4 kW electric motor, medium machines have 11–15 kW motors, and large machines are powered by 22–30 kW motors. Their horizontal operating radius is, respectively, 2.4 m, 4.4 m, and 6.5 m.

These robots can be mounted on wheeled or track chassis. The robot shown in Fig. 14.161 is equipped with a 30 kW motor and electrohydraulic drives. Its three-sectioned boom allows one to accurately position the working tool within the operating radius. Apart from the hammer, the machine can be equipped with other working fittings: concrete shears, breakers, and pulverizers. The robot is remote-controlled from a portable control panel.

Instead of the electric motor, a diesel engine or a diesel–electric drive can be used as the prime mover. The boom's fittings include: a hydraulic breaker, crushing jaws, a loader or excavator bucket, and a gripping device. Detailed information about demolition robots can be found in [14.55].

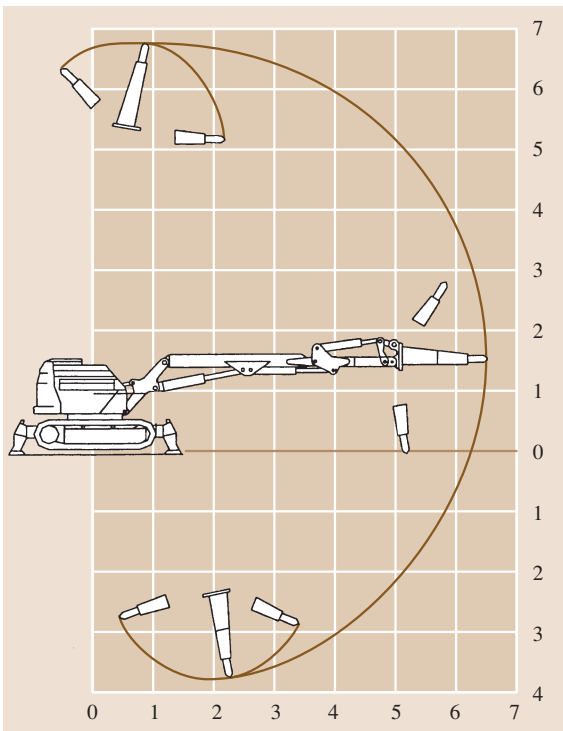


Fig. 14.161 Operating radius of robot with 30 kW motor for demolition work and material size reduction

Robots for Assessing the Technical Condition of Building Structures

Robots for checking the technical condition of building structures, referred to as inspection robots, are used for inspecting the following elements:

- Exterior wall facings
- Utility piping
- Concrete structures such as water dams and bridges
- Underwater structures

The most numerous in this group are tiled elevation inspection robots. Depending on their function, they check the adhesion of tiles to the base or check for layer corrosion. As time passes, adhesion decreases and, since the tiles may start falling off the wall, it becomes necessary to inspect the tiled walls at regular intervals.

A schematic of a robot for checking the adhesion of tiles to the building's elevation is shown in Fig. 14.162. The robot is drawn up on by a chain secured to the roof edge.

The check is conducted by continuously tapping the lining with ten small balls arranged in a row and analyzing the sounds generated. The diagnosis results, including the tapping locations, are transmitted to a computer on the ground, saved on a diskette, and represented graphically. The robot can also be used to assess the adhesion of plasters. It can operate at a rate of 700 m^2/day [14.53].

Besides tapping tiles with balls, other techniques are used to check the adhesion of tiles to the elevation, e.g., vibrating the wall and measuring the vibration characteristic. These vibration measurements are subjected to an analysis that differentiates between the frequency distributions of unstuck and adhering tiles.

Another inspection robot diagnoses layer corrosion of lining tiles. It moves on walls and is equipped with a tapping unit, a directional microphone, and an analyzer. Exfoliation is assessed on the basis of acoustic wave attenuation characteristics.

An automatic piping corrosion inspection system is intended for diagnosing the technical condition of cold and hot water systems. The inspection system can be used to check damage to 50–150 mm-diameter pipes. It measures wall thickness in a range of 1.5–9 mm by means of ultrasonic probes. A probe scanner mounted on a pipe automatically travels along the pipe, maintaining appropriate contact with it. A data processing and control unit controls the traveling and scanning move-

ments of the scanner and processes the data from an ultrasonic thickness meter [14.53].

Owing to the small diameter (150 mm) of its body and its four thrusters, a robot for the visual inspection of underwater structures can work in narrow passages, and even in pipes. A camera mounted on the slewing gear provides a wide view. The robot can work to a maximum depth of 30 m [14.53].

Depending on the user's needs, a robot for checking the technical condition of concrete structures such as bridges can be equipped with various fittings for:

- Nondestructive testing of structures
- Servicing and maintaining high-rise buildings
- Cleaning and renovating building elevations

The robot serves as a carrier for equipment and can climb vertical surfaces made of various materials. A special mechanism (with patented kinematics) with extremities in the form of vacuum suction cups [14.55] enables the robot to move on such surfaces.

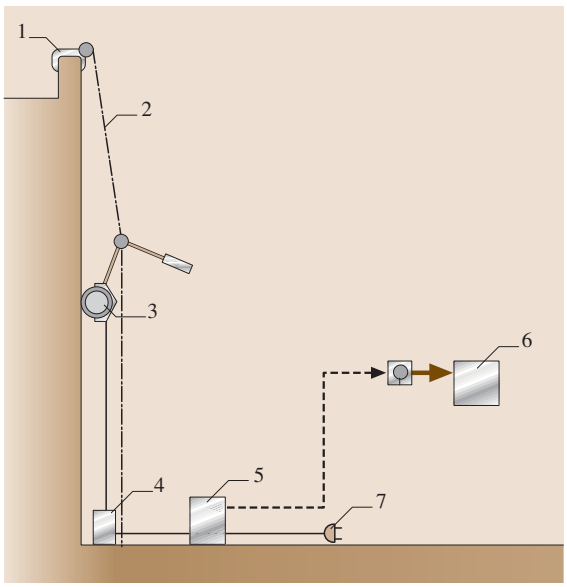


Fig. 14.162 Schematic of robot for checking adhesion of tiles to building's elevation (*1* securing robot to roof edge; *2* chain; *3* robot (hoisting gear motor and carrier of: tile adhesion diagnosing balls, analyzing circuit, and driving and communication control circuits); *4* power supply unit (performs various robot positioning functions); *5* winch controlling computer (position recording and communication control circuits); *6* output units (computer, *X*–*Y* plotter); *7* 100 V AC power supply plug)

Cleaning and Renovation of Building Elevations

Cleaning and surface-finishing tasks in buildings present a cost-effective opportunity for the application of robotics [14.61]. Several robots and automated devices for the cleaning and renovation of building elevations are available on the equipment market. Depending on the kind of elevation material (facing boards made from mineral materials, glass, etc.) appropriate renovation methods are used.

Elevations made from mineral facing boards are cleaned by blasting with an abrasive, by brushing or spraying with high-pressure water. Both the abrasive and the water are collected for reuse [14.53, 55]. Glass elevations are washed with hot water and detergents by robots equipped with sets of cylindrical and wheel brushes [14.55].

Room Air Cleanliness Monitoring

Room air cleanliness monitoring robots measure the following:

- The amount of dust in the air; in one measuring position the device collects three samples at different heights and determines the occurrence of dust divided into different diameter fractions [14.55]
- The room's environment characteristics (air flow, air pollution, temperature, and humidity) [14.55]

Cleaning of Construction Equipment

Construction equipment cleaning devices includes the following [14.62]:

- An automatic system for cleaning aluminum scaffolding boards
- An automatic washing station for cleaning the undercarriages of construction machines

The system for cleaning aluminum scaffolding boards removes concrete and dirt particles adhering to the board by means of high-pressure (40 MPa) water jets (Fig. 14.163) and ultrasonic vibration.

After washing with six high-pressure water jets the board is subjected to ultrasonic vibration. Fourteen vibrators are installed at the bottom of a large tank in which the board is immersed. In order to increase washing efficiency degassed water is used.

In the automatic washing station for cleaning the undercarriages of construction machines, washing is carried out in three stages: low-pressure water washing, high-pressure water washing, and low-pressure and

high-pressure water washing combined. The main components of the washing station are: water spraying and purifying equipment and a washed-out mud removal installation. Machine-specific automatic cleaning programs are used. All the pumps, the water purification circulation, and the removal of washed-out mud are controlled as a whole [14.53]. Force control of the robotic manipulator plays an important role in achieving successful automation of the cleaning tasks [14.63].

Surveying for Construction

Modern surveying techniques for construction are based on the global positioning system (GPS), which uses artificial satellites. The GPS has been upgraded so that it can be used in construction. The new surveying techniques are labor-saving and highly effective and do not require visual contact between the surveying stations. GPS surveying can be conducted across barriers, such as mountains or buildings, which make visual contact impossible. Moreover, surveying can be conducted in adverse weather conditions, such as fog or rain, 24 h a day, and with high accuracy.

The available surveying systems for GPS-based terrain projection use fixed datum points in combination with a mobile surveying station mounted on a truck or in the form of a remote-controlled crawler robot. Datum coordinates are transmitted to a point whose position is known [14.53, 56].

These surveying systems find application in the building of dams, roads, airports, housing estates, and so on.

Robot Implementation Issues in Construction

Construction is the single largest contributor to the national economies of most developed nations. However, the level of automation and robotization in this industry falls significantly behind that in manufacturing and many service-oriented industries [14.64]. Much of the predictions made in 1980s regarding the widespread use of robotics in construction by the beginning of the 21st century have not yet materialized. However, the core technologies and the prototypical applications of robots in a wide variety of construction tasks have been successful, as outlined in this section. Much research and development remains in order to reach the full application potential of robotics in this highly fragmented and diverse industry [14.65, 66]. New areas for future research in this domain focus on systems engineering work to redesign and re-engineer construction tasks and work sites to meet the capabilities of construction equipment. The theoretical foundations for this research were developed in the late 1980s and during the 1990s [14.67–70], but widespread support and investment from the construction industry in most countries, with the notable exception of Japan, is still mostly lacking. It is expected that the advent of applications of advanced sensor technologies and sensor networks, as well as integrated construction management systems utilizing enterprise resource planning (ERP) technologies and web-based project portals, for construction site instrumentation will contribute to significant growth of integrated, rather than single-use, fleets of robotics for application in this industry [14.69, 70].

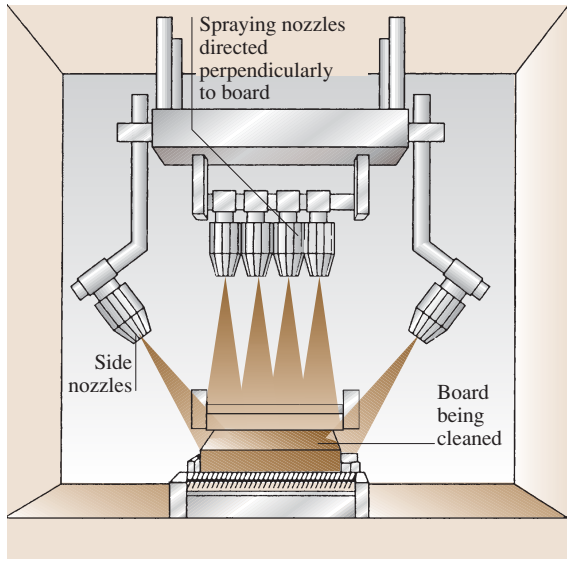


Fig. 14.163 Cleaning of aluminum scaffolding boards, stage I: cleaning with high-pressure water jets

References

14.1 O. Bachmann, H.H. Cohrs, T. Whiteman, A. Wislicki: *The Clascic Construction Series – The History of Cranes* (Giesel, Isernhagen 1997), published by KHL Int. Southfields

14.2 A. Wislicki: *The History of Excavators and Dredgers up to the Beginning of the Twentieth Century*, Editions A.T.M., Vol. 22 (Malakoff, France 1995)

14.3 ISO: *Technical Report ISO/TR 12603:1996: Building Construction Machinery and Equipment – Classification* (ISO, Geneva 1996)

14.4 Richtlinie 98/37/EG des Europaeischen Parlaments und des Rates, 22. Juni 1998 zur Angleichung der Rechts- und Verwaltungsvorschriften der Mitgliedstaaten für Maschinen (1998) ABLI.EG vom 23.07.1998. Nr. 207. p. 1, in German

14.5 F. Meier, K. Herrmann, K. Krombholz: *Einhundert Jahre für die Landtechnikindustrie* (Maschinenbauverlag, Frankfurt 1997), in German

14.6 FAO: *World reference base for soil resources* (Food and Agriculture Organization of the United Nations, Rome 1998)

14.7 ISO: *ISO 14689-1:2003: Geotechnical Investigation and Testing. Identification and Classification of Rock. Part 1: Identification and Description* (ISO, Geneva 2003)

14.8 D.G. Rossiter: *Lecture Notes Principles of Soil Classification* (International Institute for Aerospace Survey and Earth Sciences (ITC), Enschede 2001)

14.9 K.T. Renius: *Traktoren: Technik und ihre Anwendung* (BLV, München 1985)

14.10 H.-D. Kutzbach: *Allgemeine Grundlagen Ackerschlepper, Fördertechnik. Lehrbuch der Agrartechnik*, Vol. 1 (Parey, Berlin 1989), in German

14.11 W. Söhne: Druckverteilung im Boden und Bodenverformung unter Schlepperreifen, Grundl. Landtech. **5**, 49–63 (1953), in German

14.12 H. Schwanghart: 3.3 Reifen – Reifen/Bodenverhalten Tyres – Tyre/Soil-Performance. In: *Jahrbuch Agrartechnik – Yearbook Agricultural Engineering*, Vol. 16, ed. by H.J. Matthies, F. Meier (Landwirtschaftsverlag, Münster 2004) pp. 67–72, in German

14.13 D. Lemser: Radlader sind nicht nur Baumaschinen, Schüttgut **4**, 298–309 (2002)

14.14 J. Pantermöller: Funktionalität und Design bei Radladern, Tiefbau **113**(4), 237–240 (2001), WISSENSPORTAL http://www.baumaschine.de, in German

14.15 DIN: *DIN 24080: Earth-Moving Machinery* (Beuth, Berlin 1979), in German

14.16 C. Holländer: *Untersuchungen zur Beurteilung und Optimierung von Baggerhydrauliksystemen*, Fortschritt-Ber. VDI Reihe 1, Vol. 307 (VDI-Verlag, Düsseldorf 1998), in German

14.17 J. Forche: Antriebsmanagement für Hydraulikbagger, Baumaschinentechnik **26**, 33–40 (2004), in German

14.18 J. Weber, E. Lautner: Intelligente Baumaschinensteuerungen und alternative Antriebssysteme, Baumaschinentechnik 2004, Schriftenreihe der Forschungsvereinigung Bau- und Baustoffmaschinen, Vol. 26 (Frankfurt 2004) pp. 41–48, in German

14.19 G. Kunze, H. Göhrung, K. Jacob, M. Scheffler (eds.): *Baumaschinen Erdbau- und Tagebaumaschinen* (Vieweg, Braunschweig 2003), in German

14.20 Hamm AG: *Oszillation* (Hamm AG, Tirschenreuth 2004), in German

14.21 D. Lemser: Maschinen für den Straßenbau. In: *Der Elsner – Handbuch für Straßen- und Verkehrswesen*, ed. by E. Knoll (Elsner, Berlin 2003), in German

14.22 Bomag AG: *Grundlagen der Boden- und Asphaltverdichtung. Bomag Anwendungstechnik* (Bomag AG, Boppard 2002), in German

14.23 M. Buschmann, R. Grundl, H.J. Meyer: Belagfertiger mit leistungsstarker und anpassungsfähiger Technik, Tiefbau **112**(12), 772–778 (2000), in German

14.24 H.J. Meyer: *Anwendung von geodätischen Positionsmesssystemen in Straßenbaumaschinen, Baumaschinentechnik 2003*, Vol. 23 (Forschungsvereinigung Bau- und Baustoffmaschinen, Dresden 2003), in German

14.25 Wirtgen GmbH: *Slipform paver SP 500 Vario – Technical specification* (Wirtgen GmbH, Windhagen 2004)

14.26 S. Velske: *Straßenbautechnik* (Werner-Verlag, Düsseldorf 1993), in German

14.27 Wirtgen GmbH: *Cold Recycling Manual*, 2nd edn. (Wirtgen GmbH, Windhagen 2004)

14.28 C.F. Goering: *Engine and Tractor Power*, 3rd edn. (American Society Agricultural Engineers, Michigan 1992)

14.29 H. Göhlich, M. Hauck, C. von Holst: 2.5 Ride dynamics – Ride safety – Driver's place. In: *Jahrbuch Agrartechnik – Yearbook Agricultural Engineering*, Vol. 11, ed. by H.J. Matthies, F. Meier (Landwirtschaftsverlag, Münster 1999) pp. 61–69

14.30 K.T. Renius, M. Brenninger: *Jahrbuch Agrartechnik – Yearbook Agricultural Engineering 2.2*, Tractor engines and transmission, Vol. 9 (Landwirtschaftsverlag, Münster 1997) pp. 57–61

14.31 ISO: *ISO 730-1:1994: Agricultural Wheeled Tractors. Rear-Mounted Three-Point Linkage. Part 1: Categories 1, 2, 3, and 4* (ISO, Geneva 2003)

14.32 H. Auernhammer: Elektronik in Traktoren und Maschinen: Einsatzgebiete, Funktion, Entwicklungstendenzen. Vol. 2 (BLV, München 1991), in German

14.33 ISO: *ISO 11783:2000: Traktors an Machinery for Agriculture and Forestry* (ISO, Geneva 2002)

14.34 ISO: *ISO 11375:1998: Building Construction Machinery and Equipment. Terms and Definitions* (ISO, Geneva 1998)

14.35 ISO: *ISO 18650-1:2004: Building Construction Machinery and Equipment. Concrete Mixers. Part 1: Terminology and Commercial Specifications* (ISO, Geneva 2004)

14.36 ISO: *ISO 11573-1:2006: Building Construction Machinery and Equipment. Concrete pumps. Part 1: Terminology and Commercial Specification* (ISO, Geneva 1998)

14.37 ISO: *ISO 21592:2006: Building Construction Machinery and Equipment. Concrete Spraying Machines.*

Terminology and Commercial Specification (ISO, Geneva 2006)
14.38 ISO: *ISO/DIS 18651:2005: Building Construction Machinery and Equipment. Internal Vibrators for Concrete* (ISO, Geneva 2005)
14.39 ISO: *EN 12418:2000: Mansory and Stone Cutting-Off Machines for Job Site-Safety* (ISO, Geneva 2000)
14.40 ISO: *ISO 11375:1998: Building Construction Machinery and Equipment. Terms and Definitions* (ISO, Geneva 1998)
14.41 G.Y. Frenkel: *Application of Robotics and Manipulators in the Construction Industry: Construction and Progress in Science and Technology* (Znanye, Moscow 1988) p. 64, in Russian
14.42 V. Araksyan, V. Volkov: *Mechanization and Automation of Heavy and Labor-Intensive Works* (Znanye, Moscow 1985) p. 64, in Russian
14.43 G.Y. Frenkel: *Robotization of Work Processes in Construction* (Stroyizdat, Moscow 1987) p. 174, in Russian
14.44 Y.A. Vilman: *Fundamentals of Robotization in Construction* (Vysshaya Shkola, Moscow 1989) p. 271, in Russian
14.45 R. Krom: *Robots in the Building Industry* (KROM, Sassenheim 1997)
14.46 S. Singh: *The State-of-the-Art in Automation of Earthmoving* (Robotics Institute Carnegie Mellon Univ., Pittsburg 2002)
14.47 E. Budny, M. Chłosta, W. Gutkowski: *Sensitivity of the Optimum Bucket Trajectory in Controlled Excavation*, Automation in Construction (Elsevier, Amsterdam 1999) pp. 99–110
14.48 E. Budny, M. Chłosta, W. Gutkowski: *Optimal control of an excavator bucket positioning*, 19th ISARC Proc. (ISARC, Washington 2002)
14.49 E. Budny, M. Chłosta, W. Gutkowski: Load-independent control of a hydraulic excavator, Automat. Constr. **12**(3), 245–254 (2003)
14.50 E. Budny, M. Chłosta, W. Gutkowski: *A bucket discharge control for a backhoe excavator*, 21st ISARC Proc. (ISARC, Washington 2004)
14.51 P. Vähä, M. Skibniewski: Dynamic model of excavator, ASCE J. Aerosp. Eng. **6**(2), 148–158 (1993)
14.52 P. Vähä, M. Skibniewski: Cognitive force control of excavators, ASCE J. Aerosp. Eng. **6**(2), 159–166 (1993)
14.53 Council for Construction Robot Research: *Construction Robot System Catalog in Japan* (Japan Robot Association, Tokyo 1999)
14.54 M. Skibniewski, R. Kunigahalli: Chap. 17: Automation in Concrete Construction. In: *Concrete Construction Engineering Handbook* (CRC, Boca Raton 1997)
14.55 IAARC: *Robots and Automated Machines in Construction* (Int. Association for Automation and Robotics in Construction (IAARC), Watford 1998)
14.56 Fujita Corp.: *Robots for Construction* (Fujita Corp., Tokyo 2005)
14.57 PENTA OCEAN Construction Corp.: *Faces on Automatic Oriented Sheltered Building Construction* (PENTA OCEAN Construction Corp., Tokyo)
14.58 Obayashi Corp.: *Big Canopy Automation System for High-rise Reinforced Concrete Buildings*, Techn. Res. Inst. Rep., Vol. 640 (Obayashi Corp., Tokyo 2003)
14.59 J. Maeda: *Development and Application of Automated High-Rise Building Construction System*, Vol. 14 (Shimizu Tech. Res. Bull., Tokyo 1995)
14.60 M. Skibniewski, C. Hendrickson: Automation and robotics for road construction and maintenance, ASCE J. Transport. Eng. **116**(3), 261–271 (1990)
14.61 M. Skibniewski, C. Hendrickson: Analysis of robotic surface finishing work, ASCE J. Constr. Eng. Manag. **114**(1), 53–68 (1988)
14.62 M. Skibniewski: *Robotics in Civil Engineering* (Van Nostrand Reinhold, Boston 1988) p. 233
14.63 Y. Zhou, M. Skibniewski: Construction robot force control in cleaning operations, ASCE J. Aerosp. Eng. **7**(1), 33–49 (1994)
14.64 M. Skibniewski: Robot Implementation Issues for the Construction Industry. In: *Human-Robot Interaction*, ed. by M. Rahimi, W. Karwowski (Taylor Francis, New York 1992) pp. 347–366
14.65 M. Skibniewski: A framework for decision making on implementing robotics in construction, ASCE J. Comput. Civil Eng. **2**(2), 188–201 (1988)
14.66 C. Haas, M. Skibniewski, E. Budny: Robotics in civil engineering, Microcomp. Civil Eng. **10**(5), 371–381 (1995), Special Issue: Robotics in Civil Engineering
14.67 M. Skibniewski, S. Nof: A framework for programmable and flexible construction systems, Robotics Autonom. Syst. **5**, 135–150 (1989)
14.68 J. Russell, M. Skibniewski: An ergonomic analysis framework for construction tasks, Constr. Manag. Econ. **8**(3), 329–338 (1990)
14.69 J. Russell, M. Skibniewski, J. Vanegas: A framework for a construction robot fleet management system, ASCE J. Constr. Eng. Manag. **116**(3), 448–462 (1990)
14.70 M. Skibniewski, J. Russell: Construction robot fleet management system prototype, ASCE J. Comput. Civil Eng. **5**(4), 444–463 (1991)

15. Enterprise Organization and Operation

Francesco Costanzo, Yuichi Kanda, Toshiaki Kimura, Hermann Kühnle, Bruno Lisanti, Jagjit Singh Srai, Klaus-Dieter Thoben, Bernd Wilhelm, Patrick M. Williams

Organizations (derived from the Greek word *organon*, meaning *tool*) are instruments for enterprise objectives fulfilment. These objectives are to perform and produce products and services. Engineering and industrial production emphasize human-initiated, controlled, and deliberately executed combinations and transformations of resources by energy and information for the supply of market goods and products. Therefore organizations in engineering and manufacturing include the planned and purposeful action of human beings. In order to meet such objectives, formal groups of people with shared goals concerning transformation execution and output performance are configured.

Any arrangements of resources devoted to objective fulfilment constitute operations functions, or for short, operations. Technical devices can be provided to execute operations for transformation steps.

The amounts of labor involved can be coped with faster and with better quality by planned division into packages assigned to individuals for well-coordinated (repetitive) execution. For the individuals involved, operations represent tasks to be fulfilled. Combinations and syntheses of tasks and responsibilities in total constitute organization structures or parts of organizations.

In this section, the focus of our attention is on noncontractual and contractual types of collaborations among independent enterprises, pooling their core competencies to form so-called *enterprise networks*, aiming to achieve a common goal. The enterprise networks considered are composed of two or more partners collaborating under a variety of bilateral relationships [15.1].

15.1 Overview

Planning in enterprises is the definition of enterprise objectives and the anticipation of activities that are necessary to meet these objectives. In manufacturing companies, engineering focuses on transformations of materials into products and market goods by deliberate use of resources. Resource planning means the coordination of human experts, materials, technology, operations, and orders to be executed with the ultimate goal of tuning their operation. Plans (the result of planning), i. e., how all the objectives may be achieved by optimal setups and processes, may cover short-, mid-, and long-term planning horizons.

Logistics comprises all activities of planning, implementing, and controlling efficient, effective flows as well as storage of goods, services, and related information from their point of origin to their point of consumption. Seamlessly integrated logistics for the purpose of meeting customer requirements is the ultimate goal of logistics. Its achievement is generally restricted by the availability of resources, technologies, and capabilities. Recent cutting-edge research envisions logistic setups as part of a wider organizational context, as the organizational structures determine all activities and operations along the relevant value chains.

Organization and operations are studied by a number of disciplines. Relevant issues may concern technical, social, cultural, and economic areas. First scientific approaches originated in administrative, and later technical and economical, contexts as well as contributions made by organization theory. Later, efforts in sociology, psychology, political sciences, contingency theory, systems theories, and management sciences appeared.

Organizations (from the Greek *organon*, meaning tool) are instruments to achieve enterprise objectives. Enterprises and manufacturing companies emphasize organizational solutions for specific business areas, drawing results from all the disciplines and approaches mentioned above. For a given set of objectives there are most appropriate organizational setups for arranging groups of people. The clear distinction of these groups from their environments motivates the institutional meaning of the term *organization*, i. e., that they are all part of the same institutional organization. In this sense, organizations are institutions whose primary purpose is to accomplish established objectives. Rational organizational behavior is best achieved through defined rules and formal authority maintained by control and coordination.

For enterprises, the context of organizations and operations is a strategic field, as smooth, efficient operations are decisive for prosperous business development. Therefore these areas are the subject of substantial research progress and ongoing dynamic developments. Resource planning focuses on sequences of operations, which take up resources, time, space, and expertise in order to produce the intended outcome. Such sequences of operations – also called processes (from the Latin *processus*) – include all activities of analyzing, controlling, implementing, and improving, e.g., by harmonized sequences of operations (batch or flow mode) and arrangements of machines and equipment. A well-known result of elaborate process planning frequently referred to is the assembly line, based on detailed labor division and precise work flow design. All these basics of

the organization are outlined in the first section of this Chapter.

Moreover economic intentions result in objective systems favoring organizational setups for the repetition of tasks (the learning curve) by adequate design of task structures and job assignments. Individuals, facilities, and tasks are assigned in a manner consistent with the objectives, in order to obtain clear, unambiguous responsibilities, organization charts (organigrams), operations plans, and job descriptions are provided, representing the full description of a hierarchical organizational setup.

Less repetitive jobs, relying on fewer routines and therefore requiring more improvisation, require more flexible team organization principles, engaging self-organization and autonomous groups of skilled people. The teams as units may be orientated by (self-similar) subordinate objective systems that may be derived from the overall objectives.

Project organizations are very short-term structural setups for temporary enterprises, where task definitions, labor division, responsibilities, and objectives are supplied from scratch. Projects may be seen as independent organization types as well as embedded in other long-term organization structures.

Very widespread and widely applied descriptions of organizations envision companies' organizations or entire process chains as consisting of a set of linked skills and capabilities. All procedures and routines to support the efficient execution of process steps are considered to be a part of this process organization. Process organizations select and optimize alternative means to transform material or objects as well as technologies by using capabilities and competencies. Thus, evaluating, establishing, developing, shaping, and maintenance capabilities in order to ensure effective and competitive operations and processes generally represent key areas of enterprise organization. For efficient planning, checking, measuring, and correcting of the skills and competencies involved, indicators and benchmarks for best practice are particularly appropriate

Electronic data processing (EDP) has particularly triggered the development of organization and operations in enterprises. Software solutions as well as information and communication technology (ICT) implementations provide strong support for organizations as well as processes. Well-established logistics and procedures for planning and control have been made available as software packages for organization and operations management. The functions most frequently supported are accounting, enterprise resource planning (ERP), manufacturing execution (MES), and process control. Specific modeling languages and models of enterprises, organizations, and processes as well as software frameworks are used to provide formalizations and software codes. Organization and ICT may therefore generally be envisioned as complementary domains, profiting from each other's progress.

Driven by an ultimately competitive organization principle (lean production), industrial organization is actually subject to a paradigmatic shift. Fewer restrictions, faster development of markets, and the spread of technology have outdated familiar mass-production setups. Substantial efforts have been invested in overcoming the strictly functional principles of labor division in favor of dynamic organizational structures. Emphasis is placed on renewal of company culture, human-centered organization, and decentralized management. The common main feature is that human creativity and improvisation is given greater decision power. More flexible and efficient enterprise organization approaches have appeared, such as as agile manufacturing, holonic manufacturing, bionic manufacturing, and the fractal company.

Widening objectives and increasing demands for company know how, capabilities, and knowledge for engineering and manufacturing are the reasons for enterprises to loosen their organizational rules and open up their organization structures. New market opportunities, improved flexibility, quicker innovations, and better chances in local markets are the opportunities. Interorganizational structures in the form of virtual organizations, extended enterprises, or collaborative network organizations are the appropriate setups.

Aided by revolutionary ICT developments, such as the World Wide Web and global communication standards, instruments for the support of distributed organization structures are continuously improving; as equipment becomes mobile and wearable at work, individuals in industrial organizations may even work simultaneously and distributed globally on the same tasks. Communication and collaboration are now considered crucial skills in such distributed enterprise organizations. The collaborative networked organization, characterized by communication skills and totally networked structures, is emerging as an organization model for the future, in which instruments that support the collection and update of enterprise knowledge are of growing interest.

For decades the classical approach of an enterprise was to invest in the required resources and thus to plan

and realize production process by using its own resources. This has changed dramatically. Today no single enterprise is able to provide all manufacturing resources and competencies necessary for the realization of ever-changing customer demands.

There are at least three areas in which industry anticipates change and for which it must be prepared: shrinking margins, flexibility, and technology. Networked collaboration between enterprises merits greater attention. Current studies on this issue focus on important features for successful long-term collaboration (e.g., trust) as well as the management of knowledge.

As business relationships between companies and their distributed sites' complementary skills and manufacturing abilities become much closer, logistics is also affected. These changes cause the emergence of interorganizational enterprise structures, which themselves cause growth in the complexity of intra- and interenterprise logistics. Enterprise networks as a dynamic interenterprise configuration of manufacturing resources and competencies have become a promising alternative to provide required manufacturing infrastructure. Such networks include a whole range of processes in a value chain, as they do not only include supply operations of planning, sourcing, manufacturing, and logistics, but also product development activities such as research and development (R&D), innovation, product design, engineering, and customer-related activities of marketing, sales, and service.

The flow of parts through a production network has usually been preplanned by a central control system. Such central control fails in the presence of highly fluctuating demand and/or unforeseen disturbances, as is regularly the situation today. To manage such dynamic networks with low work-in-progress and short throughput times, autonomous control approaches are appropriate.

The application of autonomous control in production networks leads to a coalescence of material flow and information flow and enables every unit to manage and control its flow process autonomously.

Moreover such a strong focus on customer orientation with increasing transportation volumes results in higher supply frequencies as well as *atomization* of cargo into units. Conflicts in the objectives of logistic processes call for adaptive logistic process setups, which directly derives from changes in the production organizations.

Logistics service providers need to connect their information systems: interaction rather than integration is the means to connect. The use of agent technology and adequate organizational structures might enable these challenges to be met. Heterarchies and self-controlling units supported by multi-agent systems, intelligent load carriers, wireless communication, and ubiquitous computing are seen as the main ingredients of future logistics systems. The research work generally required in logistics planning is the transfer of methods for decentralized and autonomous control strategies that have been developed for production systems. Here too, autonomous control means decentralized coordination of intelligent logistics objects. These intelligent objects make autonomous decisions based on local information. The dynamics of such systems relies on the elements' local decision-making, generating the global behavior with newly emerging characteristics.

Since all intelligent self-controlling objects are elements in a network, information and negotiation are the results of collaboration. Therefore in logistics the greatest challenges to boosting competitiveness are concerned with organizational issues, rather than technical concerns. The design of collaboration is a fuzzy topic at best; however, one can gain insight into network design by analyzing the type of collaboration the company needs and the capabilities the company has to offer.

Bearing in mind these facts, Sect. 15.10 on enterprise collaboration and logistics places an emphasis on network and cooperation issues. Supply chain management (SCM) approaches have been considered to be useful, particularly for traditional industry where intelligent enterprise network setups are still viewed with caution. However, cutting-edge industry processes have advanced towards knowledge-driven supply network philosophies, which rely on collaboration and high-end ICT applications.

Analyzing typological issues of enterprise networks will support:

- Systematic problem analysis and solution synthesis with respect to cooperating enterprises
- A more structured view of the vast amount of possible real-life cases of industrial cooperation by characterizing these cases in a systematic way
- The analysis of cases and relevant types of industrial cooperation with respect to typical problems related to decision-making, production planning, etc.

Moreover competitive logistics includes efficient, accurate transfer of customer demands to the upstream supply partners and full interoperability between all partners. In spite of all the efforts to implement and operate technical and organizational standards such as as vendor-managed inventory, factory-gate pricing or col-

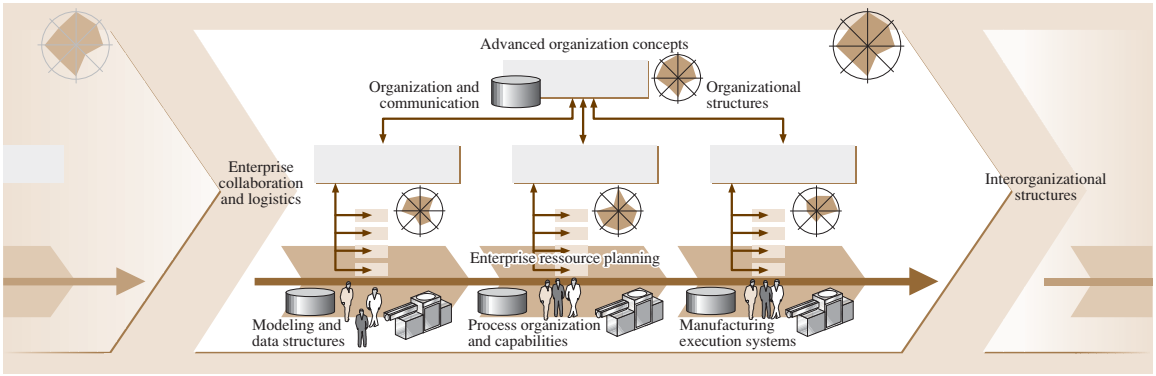


Fig. 15.1 Overview

laborative planning, forecasting, and replenishment, the key question for all partners remains: "Can I rely on my partners' information?"

In all trends underlying the dynamics of interfirm relationships, it is obvious that trust is a permanent and stable key element required to foster the implementation of any collaboration. Trust between parties involved in network activities therefore has to be considered as the key determining factor for achieving viable network outcomes. Trust, according to its various definitions, therefore plays a predominant role in the creation and development of networks. Finally, the Fig. 15.1 illustrates the main focus of the relevant topics discussed in this chapter.

15.2 Organizational Structures

15.2.1 Introduction

Enterprise organizations are large-scale social entities, where a certain level of standardization is favorable. The framework for this standardization, in which an organization defines how tasks are divided, how resources are deployed, and how departments are coordinated, is called the organizational setup. It includes:

- A set of formal tasks assigned to individuals and departments
- Formal reporting relationships, including lines of authority, decision responsibility, a number of hierarchical levels, and the span of managers' control
- Systems to ensure effective coordination of employees across units, e.g., departments

Organizations in industrial enterprises support the transformation of material and objects. Arrangements of resources devoted to these objectives constitute operations functions, or for short, operations. For the individuals involved, operations represent jobs. The job is generally the smallest entity within an organization, defining the finest degree of subdivision of tasks. Combinations and syntheses of tasks and jobs in total constitute organization structures, consisting of:

- Formal and legitimate rights of managers to make decisions, issue orders, and allocate resources to achieve organizationally desired outcomes (authorities)
- Duties to perform the tasks or activities that an employee has been assigned (responsibilities)
- People reporting and justifying task outcomes to those above them in the chain of command (accountabilities)

Manufacturing companies and industrial organizations differ from craftsmanship in terms of the complexity and larger numbers of the products they produce. Manufacturing therefore includes all human-initiated, controlled, and executed combinations of resources to provide market goods and products on larger scales. For the necessary operations and transformations, technical devices such as machines or transportation equipment are applied. Manufacturing organization and operations also includes planning, executing, and control functions for all units, performing

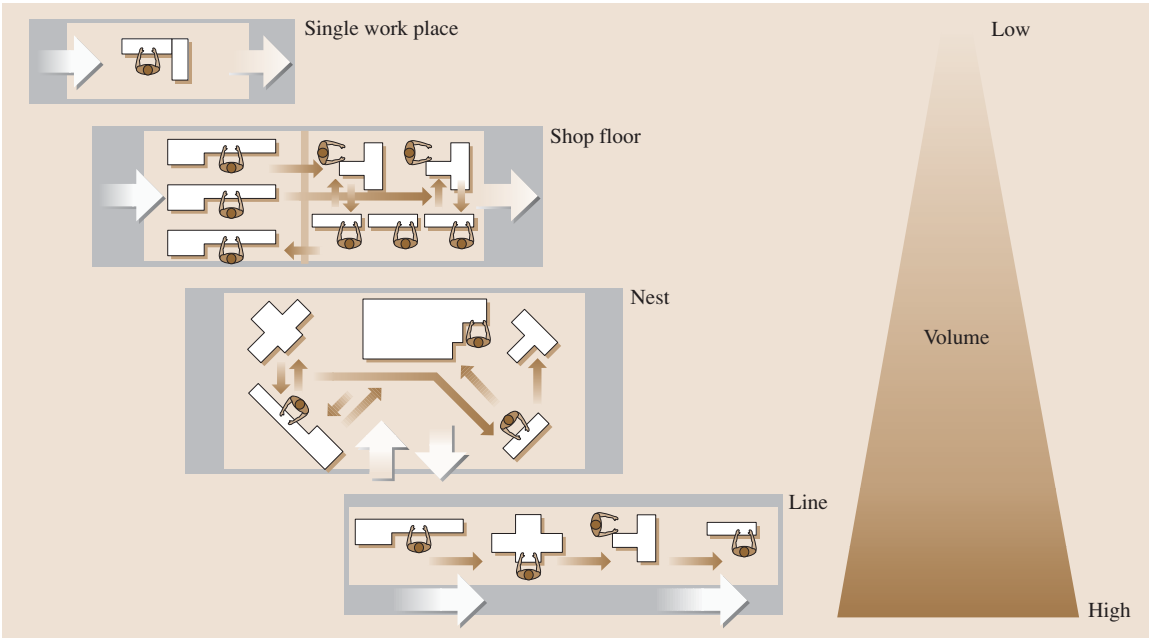


Fig. 15.2 Basic manufacturing principles

the supply, production, and maintenance of these products, goods, and services.

The basic idea of industrial organizations is to exploit the effects of multiple repetition of tasks (learning curve) resulting from labor division. This means the partition of task volumes between a number of workers, evoking the high level of specialization of individuals for a small number of tasks. Profound analysis and synthesis of all operations, functions, and tasks to be executed are the classical methods to determine the specialization requirements and the most favorable assignments of tasks to individuals, responsibilities, and technical equipment. As an information base, the definition of the sequences involved in each task, the design of the operations, etc. are established and used [15.2].

The task sequence plan prescribes all operations to be executed, e.g., assembly tasks, transportation tasks, and handling activities. The operations design specifies the operation steps to be done with more detailed descriptions (such as the machines to be used, the equipment and tools involved, machining and setup times, and qualifications needed) and order-specific information. The total information and data volumes are generally referred to as the bill of operations (BOO).

The German industry norm DIN 8580 [15.3, 4] defines and classifies important manufacturing operations.

Inputs for the definition and further specification of an operation are given by:

- Product
- Design setup
- Technology applied (technical process, time consumption)
- Volume to be produced
- Various additional parameters (material, geometry, mass, etc.)

Any sequences of operations are referred to as processes. The process definition includes all activities and instruments for analyzing, controlling, implementing, and improving industrial production under given objective settings.

Process procedures and routines of production are summarized by the term *process organization*.

Manufacturing principles describe the sum of all technological processes, designed for a volume of parts/objects, focusing on sequences of functions and geometrical arrangements (shop-floor layout) of the equipment involved (Fig. 15.2) [15.5].

Small series of a product are most frequently manufactured by applying the shop-floor principle, also referred to as the function principle, because it concentrates all machines and equipment executing identical or

Table 15.1 DIN terms and classifications for manufacturing operations

Manufacturing operation	Explanation	Examples
• Original forming (casting and molding)	• … is defined as the manufacturing of a solid body from unformed substance (metals are cast, plastics are molded)	• Extrusion, casting
• Further forming (mechanical forming)	• … is defined as a plastic change in the form of a solid body which does not change the mass or the cohesion of the body	• Bending, drawing, rolling, pressing
• Cutting	• … is manufacturing through changing the form of a solid body, thereby diminished its locally cohesion	• Turning, milling, grinding, sawing
• Joining	• … is defined as a process in which two or more bodies are linked, united, or assembled	• Welding, soldering, bonding, sticking
• Coating	• … is defined as a layer-adding process to spread a formless substance over a surface of a solid body, e.g.,for protection	• Painting, cladding
• Material property alteration	• … is the summary heading for various techniques to change material properties on the atomic level	• Tempering, hardening, quenching, nitriding

Table 15.2 Overview of the main functions of an enterprise

Procurement	Operations	Sales and distribution
Provides all goods and services necessary for the whole production process: – Raw materials – Supplies – Consumables	Comprises all kinds of operational goods and services, or is rather the realization of all intended measures for the fulfilment of operational tasks (immaterial as well as material goods). Goods and services contain immaterial and material goods	Appears at the end of the overall production process. The sale of products concludes the operational circle by initiating the reflow of cash resources
Integrating processes: provide the connections between the processing locations and fulfil the following main functions: – Transportation – Handling – Storage		
Securing processes: provide a smooth procedure for the main processes and fulfil the following main functions: – Quality assurance – Maintenance – Supply and disposal		

similar functions, e.g., milling, drilling, and coating, in the same location.

One characteristic of the flow principle is the arrangement of the equipment along the prevailing sequence of operations to be executed. The flow principle is initially derived from the product/object to be produced; therefore the term *product/object-oriented principle* is commonly used as well. Flow ideas date back to 1913, when Henry Ford introduced this principle to assemble cars, enormously increasing total effectiveness [15.6].

Production cells (production groups) concentrate all machinery and equipment covering sets of functions and technological operations necessary for the production of product or part families. Enriched by control, planning, and handling tasks, production nests have

Table 15.3 Overview of the enabling functions of an enterprise

Function	Description
Research and development	Methodical and systematic detection as well as determination and solution of causal impacts to enlarge technical knowledge
Procurement	Basically the following procurement objects are distinguished: – Operating resources – Raw materials and supplies – Staff
Production planning	Capacitated and quantitative planning of the entire production that should be realized in the upcoming period (resulting in a production programme)
Personnel planning	Providing the required employees to fulfill the following main functions: – Personnel requirements planning – Personnel procurement – Personal development – Personnel applications planning – Personnel discharge
Financial planning	Providing financial funds of any kind for the execution of operations functions
Scheduling	Supply of all the required documents for control and monitoring processes, which guarantee minimum expenditure for the manufacturing of products
Manufacturing	Production of geometrically defined solid objects by means of combining basic production factors
Assembly	Combination of individual units into modules or/and product
Marketing and sales	In addition to the name sales, other terms are applied in practice include distribution, selling, turnover, and marketing
After-sales service	Increases of customer loyalty by: – Efficient complaints management – Return of defect and surplus goods – General conditions for delivery and service (e.g., warranties)

good flexibility, high transparency, and short lead times. However, employees have to have high levels of qualifications and more skills; good levels of motivation and self-responsibility are expected.

At stationary work places a number of different functions and processes are performed on site, without having to moving the objects to be produced.

15.2.2 Enterprise: Main Functions

Generally the enterprise's organizational setup is visualized by using diagrams (organization charts) using principal functions descriptions such as procurement, operations and sales, and distribution, thereby integrating functions (e.g., transportation) and securing functions (e.g., quality assurance). These main functions represent aggregations of operations, therefore the terminology *main processes* is used as well [15.7]. The overall objective is to create value by planned input of resources. In this sense, the enterprises' main functions represent segments of value chains (compare Tables 15.2 and 15.3) [15.8].

Main functions and enabling functions are generally used for manufacturing companies. A frequently used organization chart by main functions is given in Fig. 15.3.

15.2.3 Organization and Tasks

Combinations and syntheses of tasks and responsibilities together constitute structures of organizations or parts of organizations [15.9]. For the individuals assigned, operations and functions represent tasks to be fulfilled. Task analysis decomposes comprehensive (en-

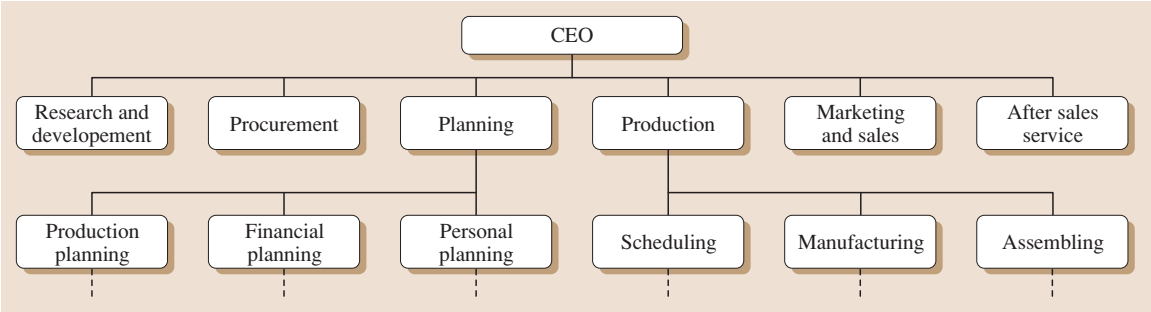


Fig. 15.3 Company main functions (overview)

terprise) tasks into subtasks and task elements. These task elements are to be assigned to task-executing units, e.g., persons, teams, and machines [15.6]. Using task synthesis, task elements are reconfigured.

Task Analysis

The overall tasks of an enterprise are decomposed by a multistage procedure into subtasks, to sub-subtasks, down to task elements, also known as elementary tasks, representing task volumes that cannot be broken down further.

The task decomposition principles may be based on the aspects of:

- Function/operation (purchase, sale, manufacturing, etc.)
- Object (products such as refrigerators and washing machines, or materials such as wood and aluminum)
- Phase (planning, implementation, control, maintenance, etc.)
- Rank (decision, operation, etc.)
- Purpose (administration, security, etc.)

Resulting task elements are linked to:

- Task-performing units within a company
- Predefined resource systems
- Available resources
- Time-dependent task volumes, execution, and fulfilment

Task Synthesis

In order to align entire processes, the analyzed subtasks and elementary tasks are clustered into sets to be assigned to units. The position (job) is the smallest unit within an organization. It includes resources and decision-making authority for resources required for the task's executions (Fig. 15.4).

Accurate job specification, demanding certain attributes and qualifications for people, can be seen as a permanent organizational challenge.

An organization involves continuous assignment of activities in the sense that labor, facility, and task elements have to be configured in such a way that the processes contributing to the overall enterprise task are supported.

A task is characterized by its

- Objectives, which should be achieved step by step
- The object to which the execution refers
- The time at which and time span during which the task is to be executed
- Responsibility (for when the task turns into an order)

In order to show clear, unambiguous responsibilities without interferences, additional instruments are provided, such as:

- An organization chart (organigram)
- Functional diagrams
- Job descriptions and documentation

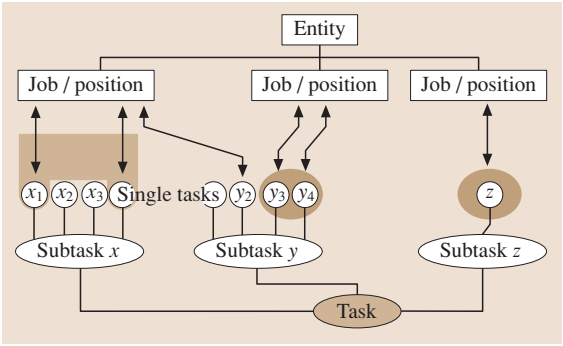


Fig. 15.4 Task analysis, task synthesis, and assignments to jobs

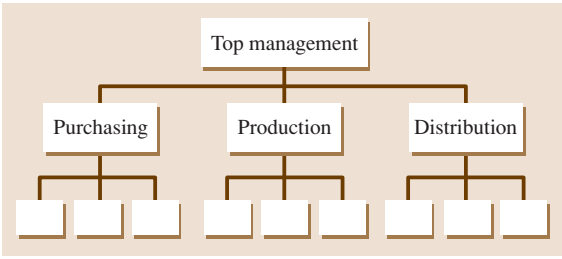


Fig. 15.5 Line organization

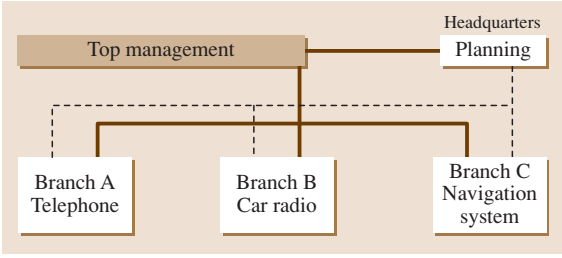


Fig. 15.6 Division organization of a consumer electronics enterprise

The collection of all these features represents a full description of an organizational setup.

15.2.4 Classical Forms of Organization

Line

Lines are derived from functions extracted from enterprises tasks, strictly hierarchically within one dimension. Orders are transmitted strictly top down, as are information and reports. Thus, an unambiguous one-dimensional information path between heads and subordinates characterizes this organization principle. Every person in the line has exactly one superior from whom they receive orders. The person is obliged to execute or pass on these orders and to report on orders' progress or status. The functional line, where the main functions of a company are structured according to the line principle, is a very common organizational setup in industry (Fig. 15.5).

Division

In this organizational form, the enterprise is decomposed into divisions, by products, product groups, services, manufacturing processes, customer groups, or regions. This object-related organizational setup allows object-oriented business units (as profit centers) to be set up within a company. Management and control is facilitated. Therefore the division structure is preferred by larger companies with diversity in production programmes (Fig. 15.6).

Matrix

The combination of the division and the line setup results in another important enterprise organizational principle: the matrix organization. Generally one hierarchy is *functional*, while another hierarchy is object-orientated, defining product responsibilities; neither of these dimensions is dominant. Matrix organizations are structures that are implemented for the manufacturing of a variety of complex technical products such as vehicles or electronic devices (Fig. 15.7).

Team

Teamwork is considered an organization structure without precise task assignment to individuals. Labor volumes resulting from rough task analysis are directly assigned to teams, where self-organizing teams take over additional tasks such as quality assurance and maintenance. By taking back the division of labor, individuals may execute more comprehensive and varying tasks. Collective responsibility for the intended

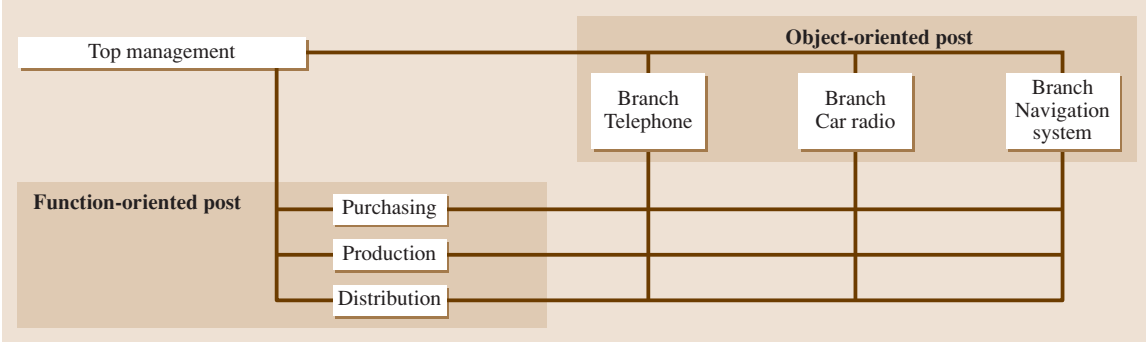


Fig. 15.7 Matrix organization of a consumer electronics enterprise

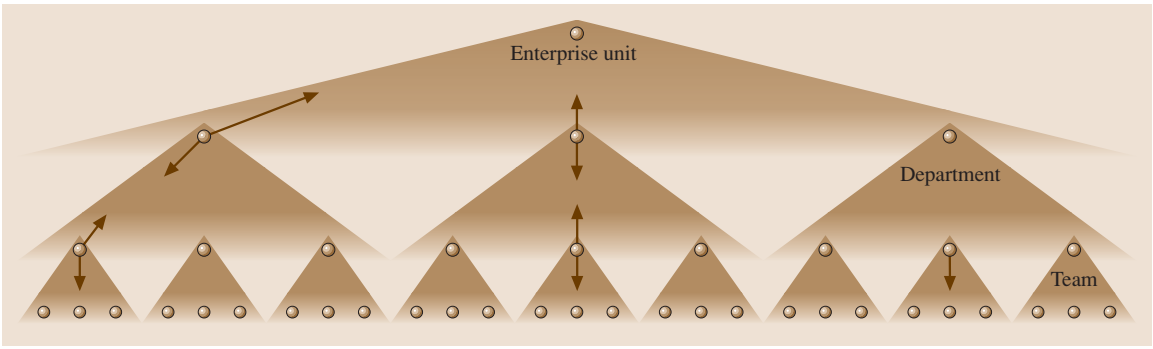


Fig. 15.8 Linked self-organizing teams [15.10]

outcome and the team members' qualifications may replace task planning and execution control. Over a longer period of time, teams implement norms and rules of behavior. Common objectives and close collaboration form the *team*. The team takes over the responsibility for the fulfillment of larger tasks (Fig. 15.8).

Teams can be classified according to their relationship to the main process areas as:

- Teams integrated in main processes
  - (Partially) autonomous teams
  - Production teams
  - Working groups
- Teams not integrated in main processes
  - Quality circles
  - Project teams

Teams and teamwork concepts offer advantages in situations where quick and flexible adaptations of structures are needed.

In cases where the enterprise tasks or task elements are subject to quick major changes, task analysis and task synthesis as well as assignments to individual responsibilities become impossible due to the lack of time and more detailed information. Teamwork or groupwork models have therefore been successfully applied in cases where speed and flexibility are required, and job enlargement (more tasks of the same type) and job enrichment (additional tasks) motivates employees [15.11].

Such process segments and the groups involved are able to reconfigure quickly (dissolve after an order, or take over a new task) and are frequently used for distinct products or for total responsibility for a project (Fig. 15.9).

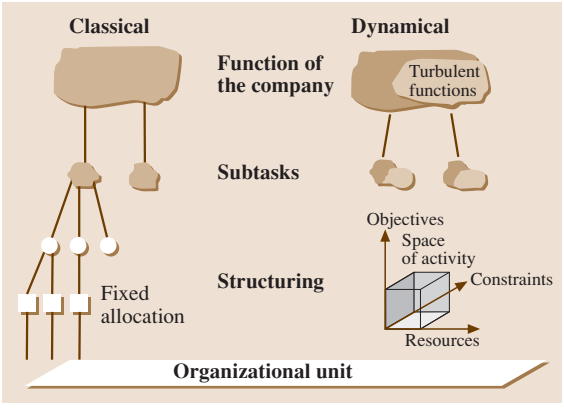


Fig. 15.9 From task assignment to process commitment

Segments and Fractals

High reconfigurability and flexibility require full commitment of all employees, which is supported by manufacturing segments. Manufacturing segments represent concentrations of machines and equipment for the complete execution of families of processes, i. e., similar processes for a set of products or product variants. Segments are an approach to combine the productivity of high volume with the flexibility of the shop floor. Segments are given a high level of autonomy (also known as *factories in the factory*), often run as cost or profit centers, in which wages are determined by output-related bonus systems [15.12].

Making even more intensive use of the advantages of self-organization, self-optimization, and dynamic team organization, the fractal organization principle decomposes the enterprise into objective-led small units. The fractal units negotiate objectives, benchmarks, and incentives, and decide on resources and working periods or operations within the assigned area. Processes are only detailed to ensure the definition of adequate unit (fractal) structures, which become totally responsible for process execution and outcomes. In combination

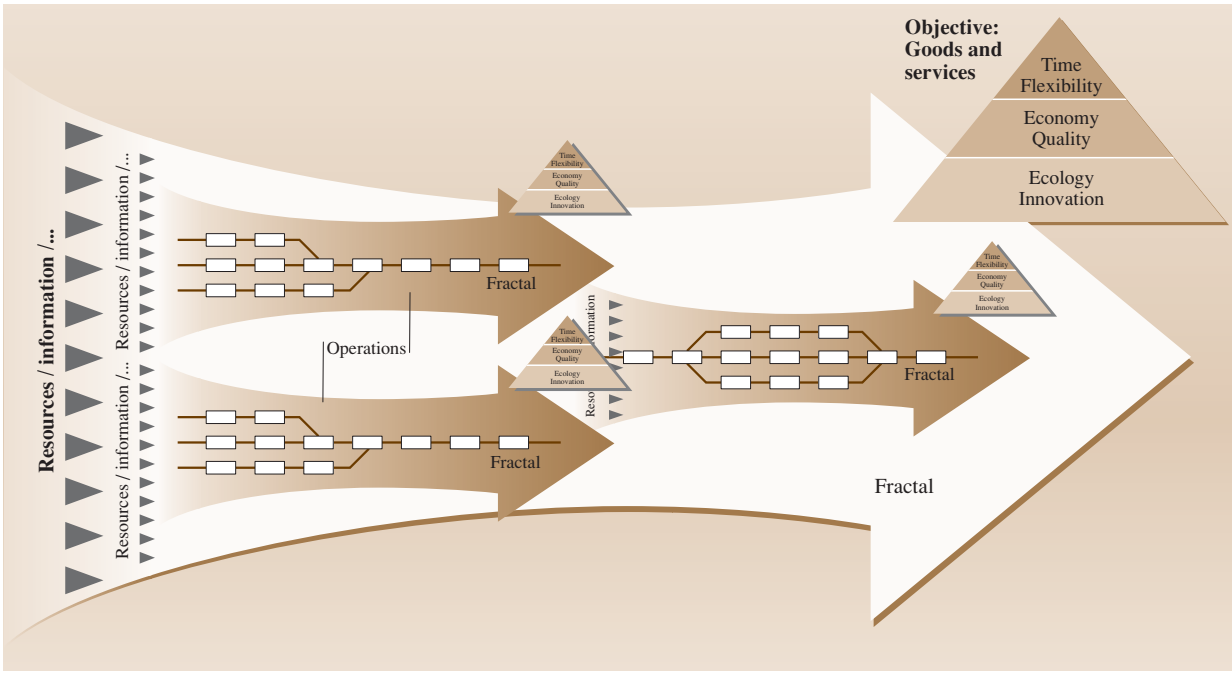


Fig. 15.10 Fractal self-similarity for objectives and processes

with incentives related to output and performance, such conditions may lead to high levels of motivation and continuous improvement [15.7]. An enterprise consisting of fractal units is referred to as a fractal company (Fig. 15.10 and Table 15.4) [15.13].

Project

Dynamic organizations are permanent organizational setups, focused on abilities to reconfigure and change over time. The setup and duration of a project organization however is limited and ends with the completion of the corresponding tasks or objects. A project is therefore a one-time task structure setup.

Project organizations are useful in various contexts. Examples are the development of a new technology, transforming organization setups, planning and execution of civil engineering objects, market introduction of new products or installations of data communication networks or software solutions in a company.

Project management includes all attributes of organizational leadership; its organization, methods, and instruments support the definition, execution, and success with respect to given objectives, including:

- Unambiguous assignments of tasks, competencies, and responsibilities
- Application of rationalizing systematic methodologies and decision support
- Ensuring information flows that support efficient and reliable decisions and controlling functions

Table 15.4 Features of fractals, the autonomous objective-led team units of an organization

Features of a fractal	Specification
Self-similarity	Fractals aim for enterprise objectives, broken down into units by negotiation
Self-organization	Operations and task assignments are done autonomously; fractals may restructure, regenerate, and dissolve
Self-optimization	The performance of all fractals is monitored for continous assessment and evaluation
Dynamics	Versatility enables coping with turbulent environment

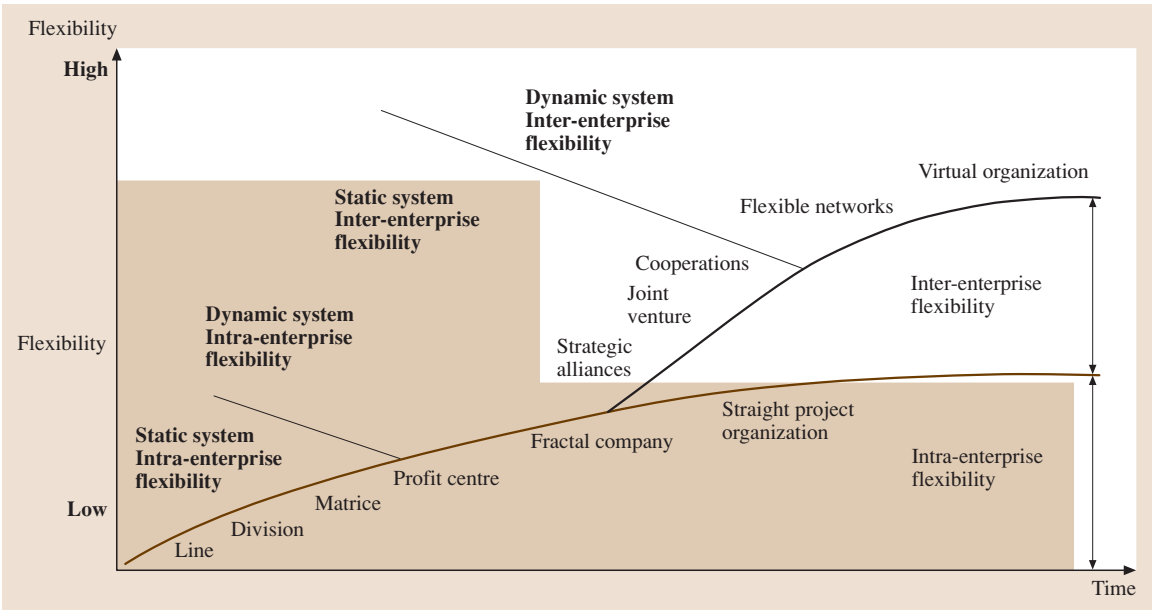


Fig. 15.11 Evolution of intra-enterprise flexibility (after [15.14])

- Building trust for collaboration within the project team
- Optimizing coordination of all units involved

Projects are defined and started in order to solve complex problems. Therefore a number of persons and units may be involved. In order to ensure that project objectives are met, projects have to be executed in a systematic way, emphasizing optimal coordination of all units involved in the corresponding phases [15.15]. A deliberate and well-prepared choice of the project leader/manager as well as the allocation of resources (front-end loading) is essential [15.16]. Project members should be assigned unambiguously to nonoverlapping tasks. Within an industrial organization there are basically three types of project implementation:

1. Genuine projects
2. Project coordinations
3. Project matrices

All of these types are characterized by a hierarchical setup. Figure 15.11 summarizes the above statements.

15.3 Process Organization, Capabilities, and Supply Networks

In general process organization defines and describes interactions between persons, equipment, and materials, executed tasks, and their fulfilment. In particular, it defines where (by which unit, by which working place) and when (in which sequence) which resources (persons, machines, etc.) transform the objects involved.

Trends point towards growing process orientation, increasing implementation of group and teamwork, and a focus on knowledge and organizational learning. Decentralized efficient units and teams reveal supplementary potential in industrial organizations and operations. Segments and product channels prove to be more powerful to cope with niche markets, in spite of possible conflicts concerning resources and management priorities.

Process organization engages flexible groups of people, including specialists across all relevant functions. Process chains describe the tuned (for a defined sequence) of process steps involving several functions. These functions are considered to be process chain elements, marked by sequence, utilized resources, op-

erations, and control. The combination of all of these function elements results in the total process, which is also referred to as the *value chain*.

Process capabilities are measures of the acceptability of variations of the process.

Usually variations in processes are interpreted as a statistical issue by engaging the characteristics of the normal probability distribution. Ambitious process capability objectives include *Six-Sigma* programs, implying 3.4 defect parts per million (ppm) and requiring elaborate instruments for process capability development.

15.3.1 The Capability Concept

Generally [15.17–19] capabilities are provided by a set of resources, networked together into a process (routine) for competitive advantage. The use of the term *capabilities*, describing firm competitiveness and business development, has a particularly rich history. The early literature on resources and capabilities has links with some of the seminal work on *rent creation* and the mechanisms to achieve this, the Ricardian perspective on *resource picking* and the Schumpeterian perspective of *capability building* [15.20]. The best usage of a firm's *bundle of resources* is discussed by *Penrose* in her original work, extending it to the area of firm services [15.21].

The resource-based view (RBV) perspective [15.17, 22] emerged in the 1980s, incorporating a broader definition of resources as a firm's *strengths and weaknesses*, and as any assets (tangible and intangible), that are semipermanently linked to the firm [15.22]. Others have suggested an all-encompassing definition of resources, such as "all assets, capabilities, organizational processes, firm attributes, information, knowledge etc." [15.23]. More recently, the resource-based view (RBV) has regained prominence under the basic premise that resources and capabilities provide the basic direction for a firm's strategy, and are the primary sources of profit [15.17, 24–27].

The term *capability* has been used interchangeably with a multitude of terms such as *strategic resources*, *organizational routines*, and *core competencies*, *meta competencies*, and these perhaps require some reflection as part of the historical context. Early RBV thinking suggested that the attributes of a firm's physical, human, and organizational capital that enable a firm to conceive and implement strategies that improve its efficiency and effectiveness are firm resources [15.22]. The separation of the capabilities element from the firm's *bundle of resources* [15.21] was captured in the seminal paper on core competencies by *Prahalad* and *Hamel* [15.28], suggesting that core competencies (essentially those capabilities that deliver competitive advantage) should provide:

- Potential access to a wide variety of markets
- A significant contribution to the perceived customer benefits of the end product
- A competence that is difficult for competitors to imitate

and that this is likely if the competence is a complex harmonization of individual technologies and production skills.

Barney [15.17] argues that the potential of *firm resources* to generate sustained competitive advantage is governed by four empirical indicators: value, rareness, imitability, and substitutability. Enterprise resources may be classified into:

- Physical capital resources (extended to include access to raw materials)
- Human capital resources: internal firm personnel, and also their relationships
- Organizational capital resources: reporting structures, planning, controlling and coordinating systems, and relationships within the firm and those with its environment

Two factors determine the value (*rent-earning potential*) deliverable from resources and capabilities [15.25]:

- Sustainability (in terms of durability, transparency, transferability, and replicability)
- Appropriability (the allocation of *rents where property rights are not fully defined*)

Various aspects of capabilities and resources in terms of value-creating potential have been developed, including: interfirm resource complementarities [15.29] (rather than similarities) that allow firms to learn new and valuable capabilities, combinative capabilities [15.30], and whether capabilities should be outcome focused, targeted at a particular desired end [15.31]. Alternative models identifying capability development approaches describe how competitive advantage may be created and sustained.

Teece et al. [15.32] recognized the need to extend the RBV into *the development and renewal of resources*

and by taking a *process* perspective involving the competencies of both internal and external resources. The focus however has been on the *integration* of external resources into the firm and the balancing of the resource mix as part of its development and renewal in a dynamically changing business environment.

Dyer and *Singh* [15.33] suggest that a firm's resources may span firm boundaries, and that these may be embedded in interfirm resources and capabilities. This *relational view* of interorganizational competitive advantage suggests an intermediate unit of analysis that sits between the *firm*-based perspectives in the extant and traditional RBV approach [15.17], and the *industry-structure*-based perspectives of *Porter* [15.26].

A parallel is drawn in the manufacturing networks arena [15.34], where there are specific capabilities of a network of manufacturing plants over and beyond those at the factory plant level. The evolutionary economics approach examines the implications of variation, selection, and retention. *Nelson* and *Winter*'s [15.35] framework is based on *routines* as the fundamental unit of analysis, and discusses their efficiency and effectiveness. Their definition of routines is indistinguishable from the RBV definition of capabilities.

Authors adopting the microeconomics approach have focused on measuring the attributes of firm resources and capabilities and correlating them with performance and, in the language of *Makadok* [15.36], form part of the *resource picking* school whilst others following the evolutionary approach form part of the *capability building* school of thought.

Capability concepts that have been developed in the 1990s [15.24] include:

- Superior rents derived from the ability to destroy and rebuild specialized inimitable resources or routines over time [15.37]
- Doubts cast on the stage theory of internationalization as new and small firms prosper (showing similarities to doubts about maturity-model stage-wise developments)
- Dynamic capabilities, i. e., specific processes that firms use to alter their resource base, as sources of competitive advantage [15.38]

The process-based perspective is given prominence by some authors [15.39], who argue that processes (and not resources) provide competitive advantage for the achievement of specific performance outcomes. This process perspective sits in between the RBV *barriers to imitation thinking* and the market forces approach (Table 15.5, based on [15.39]).

The concept of a capability lifecycle [15.40] has been introduced, incorporating the founding, development, and maturity of capabilities into several altered forms. This provides an interesting link to maturity models. Table 15.6 summarizes the key concepts, their evolution, and their implications for supply network operations.

15.3.2 Extending the Capability Concept to Processes and Supply Networks

The increasing reliance of firms on supply network partners for innovation, product replenishment, and service has complemented the traditional interfirm competition model with that of the competing supply networks model.

The evolution of the RBV therefore needs to take on board that the *organizational routines* span firm boundaries, and are not just about the interfaces within the firm, but that these network capabilities that provide competitive advantage may nevertheless follow the RBV tradition of being *rare, valuable, difficult to imitate, and not easily substitutable*.

Table 15.5 Resources, processes and performance outcomes

Operations				Markets
Resources ‘what they are’	⟷	**Processes** provide competitive advantage [15.21]	⟷	**Performance outcomes** ‘what they do’
BARRIERS TO IMITATION thinking Transform scarce resource to strategic resource (Wernerfelt 1984) Inside-out model approach (Hayes 1985, Ferdows and De Meyer 1990) Firm specific factors [15.19]				BARRIERS TO ENTRY thinking Competitive context (Anders 1984) Market influences (Hill 1984) Outside-in model approach (Miles and Snow 1984) Industry and market forces

Table 15.6 Perspectives on resources and capabilities – implications for supply network operations

Author/year	On resources	On capabilities (or routines)	Example of resources and capabilities	On competitive advantage/outcomes
Ricardo 1817	Some limited resources are inelastic	*Resource picking* (as an activity that takes place before resource acquisition)		Ricardian perspective on *resource picking* as a basis for competitive advantage
Schumpter 1950	Resource deployment and capability building	*Capability building* (an activity that takes place after resource acquisition)		Schumpterian perspective on *capability building* as a basis for competitive advantage
Penrose 1959	Tangible things (physical goods) and human resources … can be defined independent of their use (pg 24/25)	*Bundles of resources*	Resources are plant, equipment, land, FGs, WIP … and skilled and unskilled labor	Firms are *bundles of resources/services*
Nelson and Winter 1982/2002		*routines* are regular and predictable patterns of activity of the firm	Production routines, R&D, procedures, policies, … they are behavioral, inheritable, and persistant	Focuses on efficiency and effectivness of routines
Daft 1983	All assets, capabilities, organizational processes, firm attributes, information, knowledge etc. controlled by a firm that …		… enables the firm to conceive of and implement strategies that improve its efficiency and effectiveness	
Wernerfelt 1984	Any *strength or weakness* of a firm, … tangible and intangible assets which are tied semipermanently to the firm (Caves 1980)		Resources are machinery, brand names, in-house technology, efficient procedures, employment of skilled people, capital, trade contacts, etc.	Resource position barriers, attractive resources, supplementary (similar) resources, complementary resources, resource-product matrix synergies
Hayes and Wheelwright 1984	Structure/infra-structure?? (hardware/software – Slack 2005)	Structural and infrastructural capabilities of individual units (plants)	Structural elements are hardware/assets; infrastructural items are software/intangible assets	

Table 15.6 (cont.)

Author/year	On resources	On capabilities (or routines)	Example of resources and capabilities	On competitive advantage/outcomes
Prahalad and Hamel 1990	Key corporate resources include *Internal coordination and learning skills that cannot be acquired easily*	A competence that is difficult for competitors to imitate; … is likely if the competence is a complex harmonization of individual technologies and production skills	*core competences* are … coordination, learning, complex harmonization of individual technologies, and production skills	A *core competence* that is difficult for competitors to imitate
Barney 1991	Resources are the inputs, or basic units, that go into the *production* process	Capabilities (or potential competencies) of a firm are generated from a team of resources, networked together into a process (routine) for competitive advantage	Resources are assets, people skills, … capabilities are generated from a team of resources networked together into a process or routine for competitive advantage	Sustainable competitive advantage is possible if resources are valuable, rare, imperfectly imitable, and not substitutable; resource heterogeneity is a source of competitive advanatge
	Are *strengths* that firms can use to conceive of and implement their strategies	Physical capital (Williamson 1975), human capital (Becker 1964), organizational capital (Tomer 1987)		… strengths … that enable firms to conceive and implement their strategies
Grant 1991				*Rent-earning potential* from resources and capabilities governed by degree of sustainability and appropriability
Harrison et al. 1991/2001	Resource complementarity		Resource complementarity is not similarity	Resource complementarity allows firms to learn new and valuable capabilities
Barney 1992	Core competences focus on the technological and production expertise at specific points in the value chain	Capabilities are more broadly based, encompassing the entire value chain. They are visible to the customer.		Entire chain concept, visbility to the end customer

Table 15.6 (cont.)

Author/year	On resources	On capabilities (or routines)	Example of resources and capabilities	On competitive advantage/outcomes
Amit Schoenmaker 1993	*Component competence* terminology used to capture *resources*	Capabilities refer to a firm's capacity to deploy resources, usually in combination, using organizational processes, to effect a desired end	They are information-based, tangible or intangible processes that are firm specific and are developed over time through complex interactions among the firm's resources. They can abstractly be thought of as *intermediate goods* generated by the firm to prove	
Peteraf 1993				Supply inelasticity becomes a source of competitive advantage
Teece et al. 1997	Firm-specific assets that are difficult if not impossible to imitate	Dynamic capabilities as *the ability to integrate, build, and reconfigure internal and external competencies to address rapidly changing business environments*	Managerial processes and organizational (dynamic) capabilities	… the development and renewal of resources, and the integration and balancing of resources
	Assets are difficult to transfer … because of transaction costs, transfer costs, and that they may contain tacit knowledge			Internal and external resources, resource mix, internal managerial processes, resource development, and renewal
Shi and Gregory 1998		Networks of plants have additional capabilities that stem from their geographical dispersion, product and process structure/infrastructure		internal firm manufacturing network
Barney 1999	Superior deployment of internal resources is a necessary competitive requirement	Should be balanced by a broader perspective of assessing the relative capabilities of the firm and partners when formalizing the wider SC network		Realative capabilities of the firm and partners are an important factor

Table 15.6 (cont.)

Author/year	On resources	On capabilities (or routines)	Example of resources and capabilities	On competitive advantage/outcomes
Eisenhart and Martin 2000	Local abilities or competences that are fundamental to competitive advanatge	Dynamic capabilities, as specific processes that firms use to alter their resource base, as sources of competitive advantage	Resources are specialized equipment, geographic location, and human expertise.	Does not explicitly include spatial (geographic) dispersion
Winter 2003		An organizational capability is a high-level routine (or collection of routines) that, together with its implementing input flows, confers upon an organization's management a set of decision options for producing significant outputs of a particular type		
Lewis 2003	Strategic resources – as equivalent to capabilities – scarce, imperfect mobility, imperfect substitution/imitation (Wernerfelt 1984/1985)	Competence includes all resource and capability notions	Tangible (assets) versus intangible (Nanda 1996) resources	Processes (Penrose 1959, Nanda 1996) (and not resources) provide competitive advantage; resources ⇒ processes ⇒ performance outcomes (Martilla and James 1977)
Helfat and Peteraf 2003		Evolution of capabilities that encompass *operational* capabilities (collection of routines that directly contribute to production output or service) and *dynamic* capabilities (routines that build, integrate or reconfigure operational capabilities)	Parallels with PLCM. Use of terms widely used in maturity models … foundation stage, development stage, maturity stage, etc.	Introduces the concept of a *capability lifecycle*. Focus is on organizational and team capabilities (coordination) and not on the performance of individual tasks

Specifically, RBV theory promotes concepts that are, from a supply network perspective, incomplete as they are:

- Single-plant focused, i. e., their focus is on the structural/infrastructural capabilities of individual units [15.41] and they are not manufacturing network based
- Single-firm focused, i. e., not supply network oriented
- Internal assessments with few references to external network-dependent capabilities, i. e., they tend to focus on core (component) products, or internal managerial processes and organizational (dynamic) capabilities [15.32] and not multifirm (network) capabilities

The need for a more operational and network perspective of capabilities has been emphasized in the most recent literature:

Table 15.7 Supply network maturity models (adapted from *Srai* and *Gregory* [15.42])

Related cluster	Author	Maturity level 1	2	3	4	5	6
Supply network design							
SC evolution model	Stevens 1989	Baseline	Functional integration	Internal integration	External integration		
Design maturity	Fraser and Moultrie 2001	None	Partial	Formal	Culturally embedded	None	
Supply network connectivity							
e-Business development framework	Poirer 2001	Internal SC optimization	Network formation	Value chain constellation	Full network connectivity		
Supplier coordination	Hines et al. 1997	No coherent strategy	Piecemeal coordination	Systematic coordination	Network coordination		
Total network efficiency							
Quality maturity	Crosby 1979	Uncertainty	Awakening	Enlightenment	Wisdom	Certainty	
Quality management	Crosby 1996	Uncertainty	Regression	Awakening	Enlightenment	Certainty	
SC processes development and application							
ISO 9004	ISO	No formal approach	Reactive approach	Stable formal system approach	Continual improvement emphasized	Best-in-class performance	
Inventory Mgmt	Kavanaugh 2002	Become customercentric first	Acknowledge uncertainty, then exploit it	Design SCs according to product type	Find the frontier before optimizing	Pay attention to systems integration	
SCOR model	SCM 2004	Functional focus	Internal integration	External integration	Cross-enterprise collaboration		
Supplier development	Hines et al. 1997	External accreditation	Reactive problem solving	Systematic development programme	Network development		
CRM phases	Forrester 2002	Channel integration	Process redesign	Continuous optimization			

Table 15.7 (cont.)

Related cluster	Author	Maturity level 1	2	3	4	5	6
TPM – autonomous maintenance	Shirose 1992	Initial cleaning	Eliminate contam-ination sources	Create and maintain standards	Inspection	Autonomous inspection	
TPM steps	Shirose 1996	Preparation and buy-in	Formal kick-off	Execution	Establishment		
FMCG SC model	FMCG practice	Functional excellence	Service and integration	Integrated network and collaboration	Beyond known capabilities		
Software assessment	Carnegie Mellon 2002	Initial	Repeatable	Defined	Managed	Optimizing	
Product and service enhancement							
R&D effectiveness	Szakonyi 1994	Not recognized	Initial efforts	Skills	Methods	Responsibil-ities	Cont-Imp't
NPD cycle time	McGrath 1996	Informal	Functional	Cross functional	Enterprise-wide and integrated		
NPD cycle time var. 2	McGrath 1996	Informal	Functional	Project excellence	Portfolio excellence	Collaborative	
Service organization	Aerospace practice	Discrete support and services	Integrated support and services	Through-life capability	Output solutions		
Service lifecycle Mgmt	McCluskey 2004	Baseline service	Operational efficiency	Customer support excellence	Structured to grow		

- The resource-based view (RBV) in which the superior deployment of internal resources as a necessary competitive requirement should be balanced by a broader perspective of assessing the relative capabilities of the firm and partners when formalizing the wider supply chain (SC) network [15.43].
- Networks of plants have additional capabilities that stem from their geographical dispersion, and product and process structure/infrastructure [15.34].

It may be argued that taking an external-to-firm supply network perspective challenges the basic premise of RBV of *internal resources*, perhaps adopting some elements of the external environment alluded to in Porter's competitive positioning work. However, by distinguishing between closely linked supply network partners and the organizational routines and processes between them, and thus removing *ownership* as an artificial barrier to what constitutes *available resources*, maintains the philosophical links between the RBV of the firm and the supply network capabilities of closely coupled partners.

Table 15.8 Supply network capabilities

Capability		Focus and key questions
Organizational capability	Focus	Traditional focus is at the firm level, team selection
(team focus)	Q.	Do we have the right organizational structure and people?
		Strive for optimum organizational structure.
Primary capability	Focus	Operates at the functional and cross-functional level
(outcome specific)		with a focus on process development and maturity
	Q.	Do we have the right processes?
		Strive for higher levels of operational process maturity.
Metacapability	Focus	Firm strategy, firm and supply network performance with the focus on primary-capability selection
(outcome specific)	Q.	Do we have the right capability sets (profile)? Strive for optimum business model. Is resource configuration a strategic variable?

15.3.3 Application Perspectives and Maturity Models

Maturity models have been used in a number of applications, in particular in the areas of business excellence models, software development, and process improvement, including in the areas of operations management, engineering, information technology (IT), and adjacent fields. They generally use qualitative assessments or *statements* to capture stages of capability *evolution*, but may also be supported by additional descriptive accounts and by quantitative measures. However, quantitative *performance measurement* across supply networks and across industry sectors is difficult [15.44], facing the combined challenges of consistency, context, and accurate key performance indicator (KPI) assessments. Consequently the various analytical tools struggle to discriminate the best from the rest.

In addition, alignment of operational capabilities with business goals is a key dimension that is not generally covered. The supply network capability map (SN capability map) capability map attempts to address these shortcomings by taking a holistic perspective of a broad set of capabilities. Five clusters of capability are identified covering supply network design, network connectivity, network performance, enabling process excellence, and new product development.

Table 15.7 summarizes, against each of the five supply network capability clusters, some of the related maturity models used in both the academic and practice literature. These maturity models vary greatly in their level of detail but have a common approach: the notion of stagewise evolution from a basic foundation level through to a leading-edge state of process excellence.

15.3.4 Operational Process-Based Capabilities

The extension of capabilities to beyond the firm, and to supply networks in particular, makes for a process-based analysis to capability development, as well as an outcome-based assessment at the strategic level. A central issue is the development of primary capabilities developed from established processes or routines. These processes are well recognized and are often difficult to implement, but are nevertheless standard routines. They have been embraced by a whole variety of organizations. Imitation, although difficult, is possible and generally practised. They include complex processes (such as Six Sigma, total productive maintenance (TPM), total quality management (TQM), and Kanban systems), processes that have been well documented. They are regarded as transferable processes or routines that in supply network terms exist within and between functions, and across firms.

Metacapabilities, on the other hand, or higher-order capabilities, involve selective resource picking, trade-off approaches to arrive at a particular set of primary capabilities (a particular capability profile). The selection process seeks to achieve a particular outcome or alignment to the business model, and thereby provides for a unique metalevel capability that is generally far less well understood. These metacapabilities are company (or supply network) specific, customer rel-

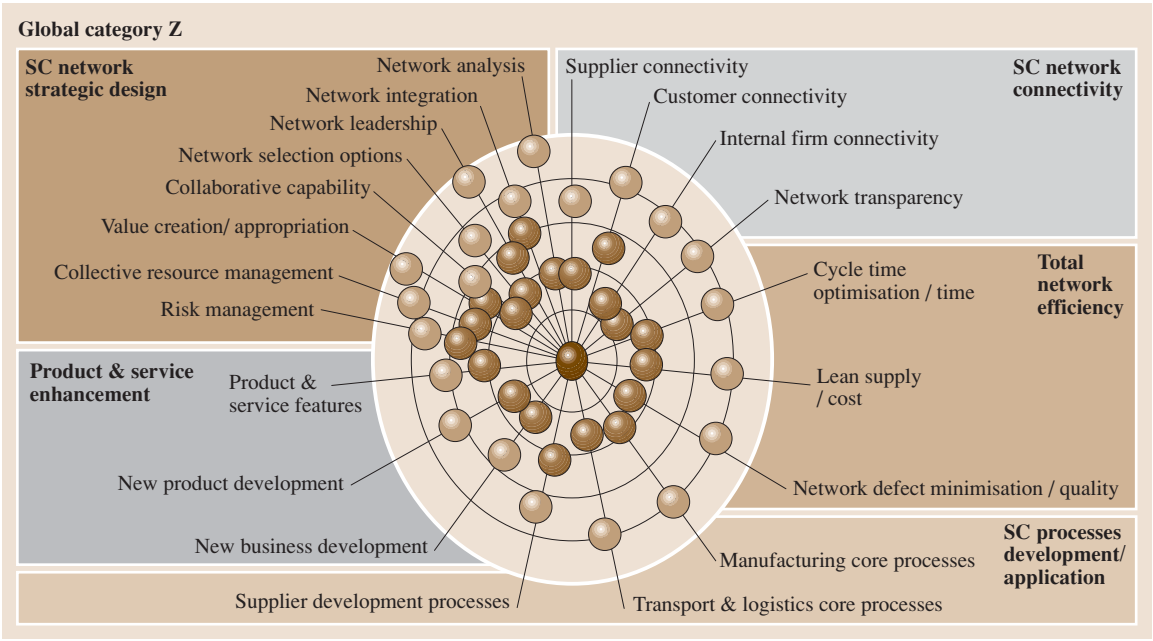


Fig. 15.12 Supply network capability map, actual versus business strategy (copyright [15.45])

evant, and business aligned. They have greater path dependency, team-dependent characteristics, and target-specific outcomes at the firm and network level. In theory, there are alternative methods for achieving similar outcomes by developing alternative combinations of primary capabilities.

Both primary and metacapabilities may be regarded as dynamic capabilities, as both evolve with time. The concept of dynamic capabilities working on operational capabilities [15.32] is not challenged in this view, and the concept of higher-order capabilities is indeed supported. However, rather than enhancing lower-level capabilities, the focus of higher-order capabilities is on overall coherency and fit with the business model, a process of selection and design for the *competitive advantage* or *outcome* that is being targeted.

In the case of primary capabilities, they mature through various stages of development, and their evolution (although difficult to achieve) is predictable. The stages of evolution in industry are captured by maturity models with well-developed processes for progressing through the various stages (compare Table 15.8).

With metacapabilities we introduce the *business model* or *outcome* approach, involving selective development of key primary capabilities to selected levels of process maturity. The overall capability profile becomes unique. Its development is less predictable and is much more difficult to imitate. Indeed the desire to imitate may be less, recognizing heterogeneity and path dependency factors, whilst the outcome will be a major focus of competitors.

15.3.5 The Supply Network Capability Map

The development of the SN capability map using the maturity model approach included the following design elements:

- The technique must maintain the business strategy/model relevance of the capability assessed.
- The definitions used should be broad and generic to have cross-sector validity.
- Increasing maturity is not necessarily valuable, as maturity is not an end in itself; rather an assessment is made of its alignment to business needs and the need to ensure a positive cost–benefit analysis.

In order to maintain a light-touch approach, five levels of maturity were used, each including description of key processes, level characteristics, systems deployed, performance standards targeted, and main activities, with only broad descriptions and limited detail on actual measurement and scoring (Fig. 15.12). The framework model, depicted as a radar chart, shows the primary supply network capability dimensions and their related

Cluster summary

Maturity level	1	2	3	4	5
Supply network strategic design	Accidental / initial	Repeatable	Defined	Managed	Mastered / optimised
Supply network connectivity	No coherent strategy	Piecemeal coordination	Systematic coordination	Network coordination	Cross enterprise alignment & collaboration
Total network efficiency	Baseline	Functional integration	Internal integration	External integration	Cross enterprise collaboration (industry leader)
SC processes developement / application	Baseline	(Reactive) problem solving	Systematic development programme	Network development	Cross enterprise collaboration
Product & service enhancement	Informal	Functional / formal	Project excellence	Portfolio excellence	Collaborative

Key: Achieved level | Current maturity level | Business SC target | Higher level of maturity

Fig. 15.13 Summary output: SN capability assessment

clusters. The five concentric circles represent the scale limits of one (inner) to five (outer).

A simpler summary table format is shown in Fig. 15.13. This has been found to be useful with regard to being more selective for areas in which to excel, moving beyond cost-based discussions with business colleagues, and the need to balance resources between maintaining continuous improvement activities and targeting new opportunities.

The model provides a holistic visualization of the network capabilities identified in the literature, and enables comparison with business strategy requirements and actual performance, allows consideration of trade-offs in supply network capabilities, and demonstrates the interplay of capability dimensions. The capability improvement task is itself strengthened using more process-based performance goals. The approach generates a capabilities profile and a gap analysis for a firm's supply network. The assessments are based on levels of process maturity, excellence models, and performance benchmarks. Opportunities and trade-offs can be assessed using processes that have both cross-sector relevance and practice history.

15.4 Modeling and Data Structures

15.4.1 Introduction

Efficient execution of operations and processes within industrial settings usually requires relevant information for all involved activities, providing instructions for how and when each activity, operation, or task must be executed for both organizational entities and the individuals involved.

In industrial production the relevant information will specify products (bill of materials, BOM) and processes (bill of operation, BOO). Based on these, production processes and the involved data processing are engineered and executed, often in an automated or semi-automated fashion.

Within the engineering as well as the execution of production processes electronic data processing (EDP) support is one of the key drivers for manufacturing and process organization developments. Engineering therefore increasingly focuses on formal design and analysis of enterprise organization structures and processes using a formalized representation of the information relevant within engineering and execution; for example, products may be embedded into an information flow from supply, through design, manufacturing, assembly, to distribution and sales. Methods of enterprise modeling support these complex tasks.

All main functions of an enterprise covering the application of relevant information, as men-

tioned above, may implement computer support through computer-aided design (CAD), computer-aided manufacturing (CAM), computer-aided process planning (CAPP), or production planning and control (PPC). Computer-integrated manufacturing (CIM) is the sum of the production systems information flows, managing all areas through central databases.

EDP has not only contributed to the support of organizations and processes, it has made technologies and information flows more *intelligent*, e.g., by integrating advances in information and communication technology (ICT) into the application/processing of relevant information. Examples are computer numerical control (CNC), net communication technologies such as field busses or industrial ethernet, advanced software design concepts such as object orientation and agent-oriented programming, and ontology-based contract protocols, to mention only some of the highlights of the last decade.

Interoperability requirements and the variety of ICT solutions, offered by increasingly modular setups of devices and software programs, highlight the importance of technical, ICT, and organizational standards, elaborated by national, European, and international institutions, engineering associations, and interest groups (e.g., the Institute of Electrical and Electronics Engineers (IEEE), VDI, ISO, and DIN).

To support the engineering and execution of operations and processes within industrial settings using EDP based on advances in ICT it is necessary to provide the relevant information in an electronically processable manner. Therefore, the relevant information has to be integrated into model-based structures.

These models separate the relevant information from the rest of the information available within industrial settings and structure it in a manner that abides by the requirements of EDP. They enables structured information collection, representation, and processing at all levels and for all purposes within industrial settings. Hence, models are an essential part of the successful application of EDP within industrial settings.

One important field of application of EDP-based information processing is the engineering and execution of organizational structures within industrial settings, i. e., the engineering and execution of business and business-like manufacturing processes.

In the following section we will discuss this field of interest with respect to information modeling-related terms, methods, tools, and models.

15.4.2 Definitions

The starting point of the consideration of modeling within the field of organizational structures within industrial settings is the term system *enterprise*. This can be described as a set of entities and a set of relations between these elements [15.46].

A *system* is defined as a *structured formation*, which is delimited or assumed to be delimited from the environment. It consists of a set of *elements* (parts) which affect each other due to fixed *relations*.

According to [15.47], there are various views to analyze systems thoroughly:

- The *functional view*, which understands system as an input/output system (I/O system) whereby the transformation of input data to output data describes the function of the system
- The *structural view*, which focuses on elements and relations between these elements
- The *hierarchical view*, which focuses on the fact that a system might be composed by a set of multiple subsystems and belong to a supersystem

According to [15.48] and [15.46] an enterprise can be understood as a system that is:

- Real (physical, observable – not theoretical)
- Open (relations to environment exist)
- Complex (large and continuously changing number of elements and relations)
- Probabilistic (prediction of future behavior has a certain probability)
- Artificial (created by humans)
- Dynamic (properties of elements and relations are variable over time)

A *model* of a system is a constructed, simply changeable, and easily understandable system that represents a hardly changeable, complex system regarding a certain problem or aspect. Therefore, only the corresponding relevant elements are mapped into the model. In general, models will only represent a subset of all the information that a system contains. Therefore, they are only projections of the original system to model systems that can be used to answer specific questions about the structure and behavior of the original system.

Various types of models can be distinguished. This distinction can be based on the set of questions that can be answered by a certain type of model as well as the viewpoint of the original system that they represent.

The first and highest-level distinction is related to qualitative or quantitative modeling. *Qualitative models* describe elements, relations, and properties *verbally*, i. e., they represent only dependencies among elements, relations, and properties without describing the level of dependency and the strength of the effects of these dependencies. In contrast to this, *quantitative models* express elements, relations, and properties by using *numbers*, i. e., they describe the strength of these dependencies among elements, relations, and properties.

Another high-level distinction between model types is the goal of model design, which distinguishes between models for description, explanation, and decision [15.49]:

- Descriptive models enable a black-box analysis of a system by providing a description of its overall behavior (data analysis).
- Explanative models enable an impact analysis of the relations between elements of a system.
- Decision models enable the evaluation of certain actions against the goals and constraints of a system.

Independent of the intended application domain of models or the set of questions to be answered by the model analysis the process of mapping the overall system information to the model system by selecting relevant information has to be well structured and to reflect the intended use of the model. This mapping process is called modeling.

Modeling is the creation of a model to describe the *structural design* of a system (*system synthesis*) and to generate methods (*algorithms*) for the evaluation of interesting system properties (*system analysis*).

The modeling process is generally iterative (i. e., a cyclic method with individual steps) consisting of:

- The definition of goals and system borders
- The definition of views for structuring of system
- The identification and modeling of elements and properties
- The identification and modeling of outer relations (relations to the environment)
- The identification and modeling of inner relations (between elements)
- The validation of system behavior
- The analysis of system behavior

Within the modeling process relevant information is collected based on the real system and modeling goals are defined, after which the modeling steps take place.

Once the model has been developed it has to be, on the one hand, evaluated and analyzed to validate or verify its correctness with respect to the intended viewpoint of the original system and, on the other hand, analyzed and interpreted to answer the addressed questions.

Based on this general modeling process, to set up *complex models* certain modeling methods have been introduced and have been proven to be useful. According to [15.50], a *modeling method* includes a modeling language (the syntax for model entities) as well as the rules for model design (the semantics of the model entities). A *modeling language* is characterized by a graphical notation and the description of the notation syntax. The rules for model design are described by the modeling steps, their fixed execution order and dependencies, and the results reached by the individual modeling steps (Fig. 15.14).

Every *modeling method* is based on a proper set of rules and expressions, i. e., model design rules and steps and modeling language constructs, for:

- Model design and model adaptation
- Recognition of reality and its design (the author's view)
- Decrease of risk and cost reduction (the user's view)

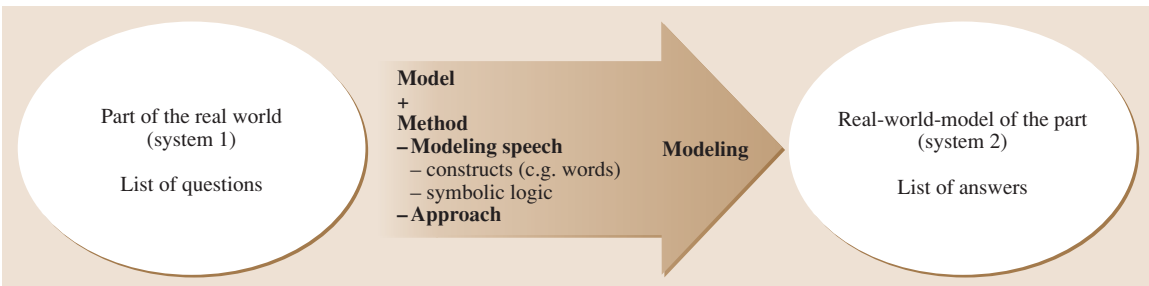


Fig. 15.14 Distinction between model, method, and modeling language

Modeling languages and modeling methods provide the basis for modeling technologies. While modeling languages define the syntax of models, modeling methods define their semantics.

To increase the applicability of modeling languages and modeling methods and reduce modeling risks and costs, reference models are used. A *reference model* provides the structures, properties, relation, and behavior of objects of an application domain in a general and usable form, supporting the creation of *specific models* by adaptation. Usually, reference models consist of various submodels or views, describing the modeled facts in varying levels of detail, and varying points of view on the original system [15.51].

Often-used viewpoints are the static structure of a system and the behavior of the execution of the system to be modeled.

The benefits achieved due to usage of reference models are:

- The recognition of weak points and the resulting improvements
- The documentation of existing workflows
- The standardization of terminology

A special type of a reference model is an architecture. An *architecture* defines the basic structure of a system resulting from a set of complementary and superimposing substructures and partially complementary viewpoints on the overall system. The architecture determines the structure of a system using two layers. On the one hand, the architecture establishes the model of the regarding system. Thus, it defines the object layer. On the other hand, the architecture defines rules that have to be considered during the development of the system [15.52].

Modeling architectures, as the most powerful but also most application-domain-oriented technology, form the top level of modeling technologies. They provide sets of models with fixed modeling syntax as well as model semantics related to the application domain. Thereby, they enable fast and efficient model system design and analysis by adapting inherent reference models to the original modeled system.

15.4.3 Guidelines of Modeling (GoM)

As mentioned above the modeling process requires a well-structured organization and well-defined inputs and outputs to reach the modeling objectives of providing a view of a system to answer dedicated sets of questions.

A framework to ensure the objectivity and correctness of modeling is provided by *guidelines of modeling* (GoM), the goal of which is to provide design recommendations for modeling to improve the quality of models beyond the matching of syntactic rules [15.51]. The guidelines of modeling (GoM) are given as follows:

- General principle of correctness:
  - Correct reflection of the mapped issue (semantically, the described structure/described behavior; syntactical, the consideration of notation rules)
- General principle of relevance:
  - Documentation of the relevant issues regarding the respective view
  - No mapping of irrelevant information
- General principle of efficiency:
  - Modeling activities shall be performed with a suitable cost–value ratio, e.g., through the use of reference models and support of reuse
- General principle of clearness:
  - Structure
  - Clarity
  - Readability (intuitive)
- General principle of comparability:
  - Comprehensive application of modeling rules
  - Goal: consolidation of independently created (sub)models
- General principle of systematic setup:
  - Well-defined interfaces to corresponding models (e.g., input data of the process model/reference to data model)

Applications of models are constrained by the method chosen for modeling. Models and methods cannot be combined arbitrarily. The following applications are available as modeling methods:

- Description of functional structures (aspects of operations in production systems)
- Mapping of functions and specification of organizations (system analysis)

15.4.4 Important Models and Methods

In this subsection examples of architectures, reference models, as well as modeling languages and methods relevant within to the field of organization modeling will be described. Thereby, the main characteristics of the different levels of model system design supported by them will be emphasized.

Modeling Architectures

Modeling architectures provide the possibility to model the original system by describing different points of view within different models, all of which are equipped with a predefined syntax and semantics of the model entities.

In the following the three architectures relevant for modeling of organizations and organization structures are described: the architecture of integrated information systems (ARIS), the generalized enterprise reference model architecture and methodology (GERAM), and the computer-integrated manufacturing open system architecture (CIMOSA).

ARIS. The architecture of integrated information systems (ARIS) is a modeling framework of methods for systematically developing enterprise application systems. ARIS is designed to reduce the complexity of enterprises.

This is done by separation into five specification views: the organization, the data, the control, the function, and the performance. This differentiation allows the application of specific methods for each view without having to consider the relationships to other defined views (Fig. 15.15).

Additionally, according to the general lifecycle concept, specification methods for information systems are used. The layers of specific concepts, data-processing concepts, and implementation describe an economical issue from the emergence of a problem to technical implementation. Through specification views and layers, ARIS defines an architecture in which selected proper (known) modeling methods are combined to a holistic method of business modeling [15.54].

GERAM. The generalized enterprise reference model architecture and methodology (GERAM) is understood as a *toolkit concept* for business process planning and design over the whole lifecycle of an organization. According to the IFIP-IFAC (International Federation for Information Processing, International Federation for Automatic Control) Task Force and ISO/DIS (Draft International Standard) 15704 [15.55,56] GERAM provides a pragmatic approach for a holistic framework to describe business processes due to different business development processes (e.g., founding, reorganization, fusion, re-engineering, setup of a virtual enterprise or supply-chain integration).

In this sense, all necessary elements are described; standards for tools and methods are set for useful accomplishment of integration and changing processes.

The basic reference model architecture provides an analysis and modeling framework based on the lifecycle concept that identifies three dimensions for defining the scope and content of enterprise modeling:

- Lifecycle dimension, providing a controlled modeling process of enterprise entities according to the lifecycle activities

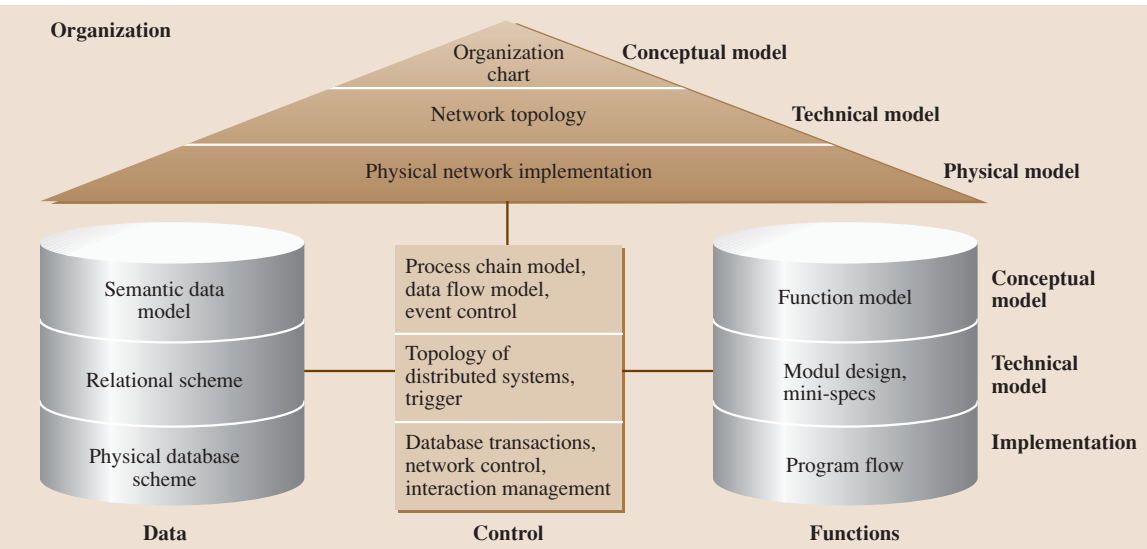


Fig. 15.15 View-concept of ARIS (after [15.53])

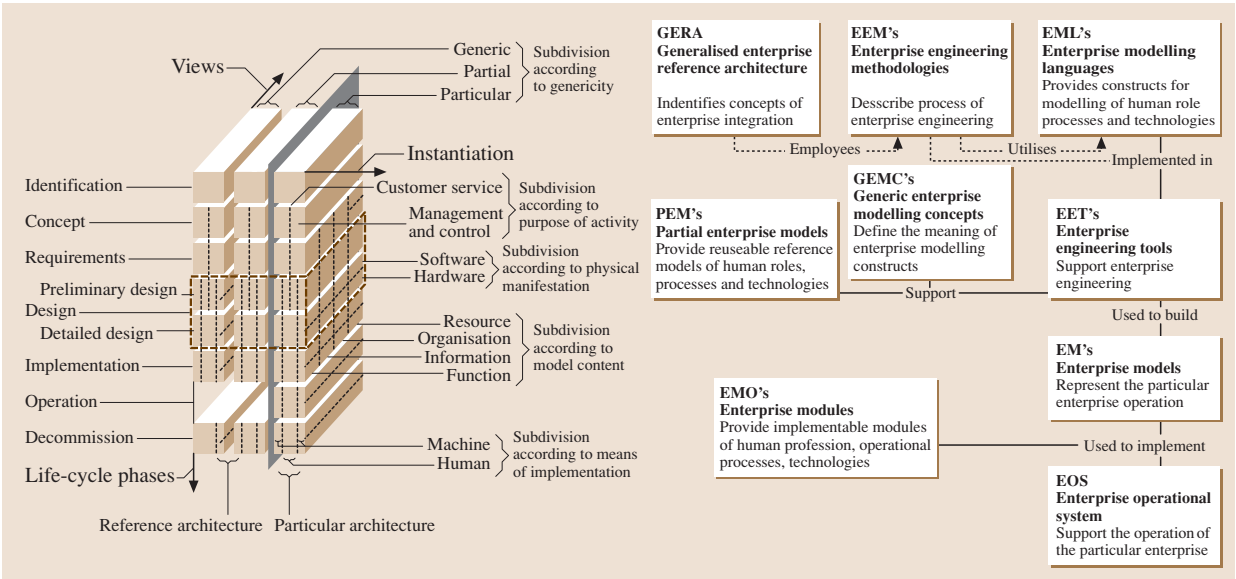


Fig. 15.16 GERA modelling framework and methodology components (after [15.55])

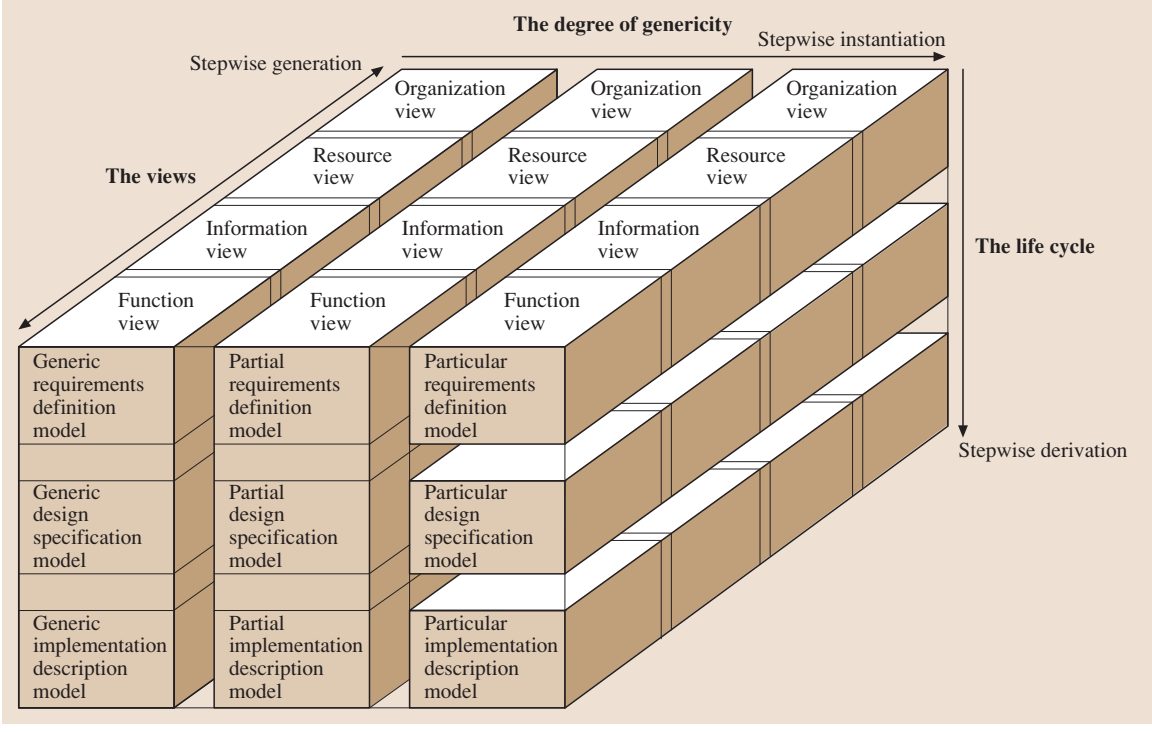


Fig. 15.17 CIMOSA architectural framework [15.57]

- Genericity dimension, providing a controlled particularization (instantiation) process from generic and partial to particular
- View dimension, providing controlled visualization of specific views of the enterprise entity

Figure 15.16 shows the three-dimensional structure identified above, which represents this modeling framework and the components of the methodology.

GERAM comprises several reference architectures for enterprise modeling as it integrates the content and aspects of GRAI integrated methodology (GIM), CIMOSA, and purdue enterprise reference architecture (PERA).

CIMOSA. The computer-integrated manufacturing open system architecture (CIMOSA) provides a process-oriented holistic procedure for enterprise integration for companies following the CIM philosophy. The variety of models and the necessary interaction between models as well as the transfer from planning to operative models require a model for modeling. To achieve this, the CIMOSA concept uses a three-dimensional frame to support the mapping of different aspects and views (Fig. 15.17).

The derivation dimension arranges the models in a sequence according to project phases: requirement definition, design specification, and implementation description. Instantiation arranges models according to increasing degree of abstraction or decreasing concretion. Starting from individual or particular models of business, it generalizes according to similar applications (e.g., industry sectors) to achieve partial models. The next step results in generic constructs and basic atoms. The opposite sequence is used for implementation. The generation dimension arranges complementary models which emerge from distinct technical disciplines (including their specific terminologies) and with different views on the factory.

The modeling is done in three phases: requirement definition, derivation of design specifications, and final implementation. Thereby, four views are introduced: function, information, resource and organization view [15.59].

CIMOSA modeling in an assembly plant analysis is illustrated in Fig. 15.18. This application includes the global modeling of processes for ordering parts and products as well as the material flow control in the manufacturing–assembly–warehouse environment. Modeling the different processes in the manufacturing domain allows the simulation of different business scenarios and the analysis and evaluation of the results in terms of throughput and turnaround time.

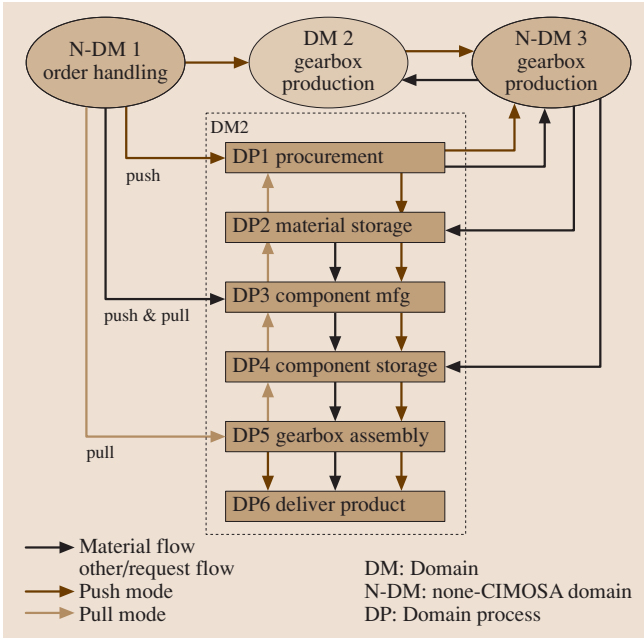


Fig. 15.18 CIMOSA model of an assembly plant analysis [15.58]

The Six-Layer Model. A specific CIMOSA derivate for description as well as implementation of organizations is used for fractal self-similar objective and structure break down. The cooperation between the fractals in a fractal company (Sect. 15.1) takes place on several levels, which can be seen and handled as a whole but which are separate from each other (Fig. 15.10) [15.60].

A company is defined as a social-technical system with a large number of subsystems enclosing process chains. To research and use these complex and nonlinear systems a model that allows a usable and complexity-reducing view and which otherwise avoided a particulate view on single aspects is required.

The model's basis is a horizontal integration of the company's functions. The results are different aspects (usable when isolated), without deleted the integration view of the whole company's processes [15.60]. Figure 15.19 provides an overview of the six-layer model.

There is no master plan to implement the basic principles of the fractal company by using the six-layer model. For successful implementation it is essential not to stay focused on a single aspect, but rather to use an integrated view of the subsystems. In this way it is possible to generate coherent solutions over all levels.

Reference Models

Reference models provide a view of a system based on a predefined syntax and semantics of the model elements.

In the following, three types of reference models relevant for modeling of organizations and manufacturing systems are described: operations modeling, event-driven process chains (EPC), and the entity relationship model (ERM).

Operations Modeling. A *process* consists of a number of continuous actions – operations, transports, and idle states – on an eventually manufactured part/object. If the latter consists of matter it is called material (in the broader sense).

It is usual to design a temporally limited (improvisation) or unlimited (organization) execution of actions if equal or similar actions are repeated. This defines (among other results) the path along which the objects will pass during the corresponding actions. The objects using this path form a *material flow*. The processes are characterized by the following properties:

- Contents: the type and size of objects or flow entities
- Strength: the number of objects per time unit
- Speed: the distance traveled per time unit
- Intervals: the constant or variable temporal distance between objects (for constant flow)
- Limitation: the path, channel, or pipeline, including reservoirs involved
- Direction: related to channel limitation
- Length: measured as a spatial distance or the number of process states
- Layout: the spatial configuration
- Consistency: duration of flow existence
- Significant points

Processes and operations may be represented in various notations, e.g., using operation sequence diagrams based on definitions provide by the Association of German Engineers (*Verein Deutscher Ingenieure*, VDI) or the American Society of Mechanical Engineers (ASME).

Following symbols mark the different actions (compare Fig. 15.20) [15.52].

Operations:

- ◯ Work: Object gets closer to desired final state.
- □ Check: Measurement and comparison with reference, recognition of differences.

Movements:

- ⇒ Transport: Self-contained action of location changing.
- ◯ Handle: and put, place, load and unload, move, accompaniment (auxiliary time) in conjunction with operation, transport, or storage.

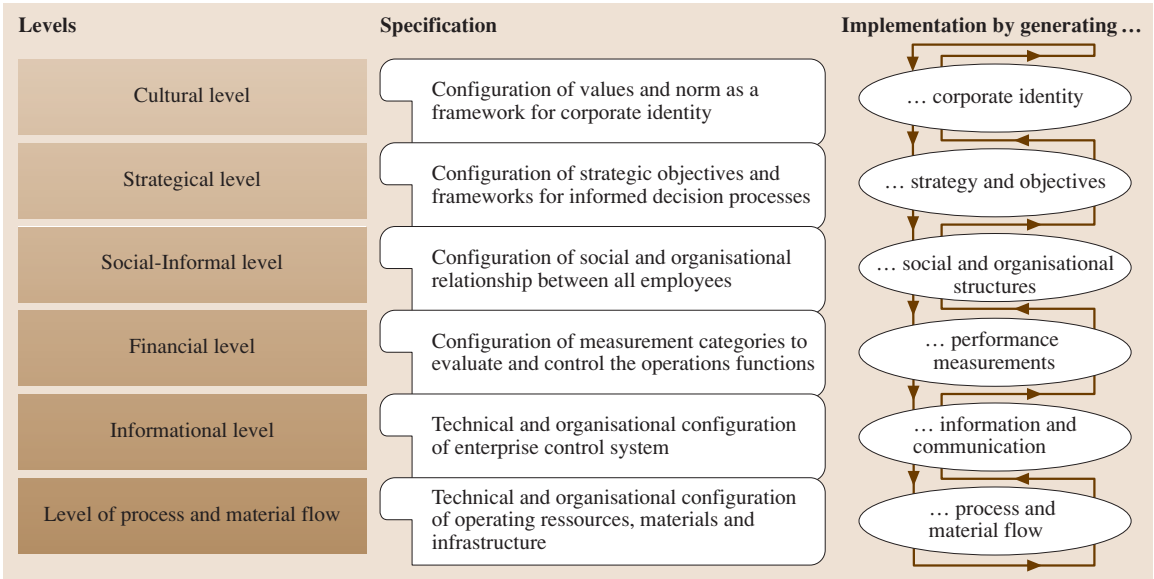


Fig. 15.19 The enterprise six-layer model (after [15.61])

Idle states:

- ▽ Storage: Intentional interrupt of material flow.
- D Delay (wait): Unintentional interruption of material flow: malfunction, stagnation, jamming, waiting at construction, or destruction of transport units.

To simulate discrete event processes, the following types of model elements can be distinguished [15.62] (compare also Fig. 15.21):

- Server (all time-consuming system elements, characterized by capacity, interarrival time, and throughput time)
- Model border points (sources, income of movable elements from the environment; sinks, emission of movable elements to the environment; characterized by the interemission time)
- Control points (unification or splitting of material flows, characterized by function and interemission time)
- Connectors (connection elements, characterized by direction)

Flows and processes can be sketched. In order to explore and forecast process design decisions richer simulation models are applied.

Before a simulation can be started, the structure and logic of the system (operation control) has to be modeled (i. e., presented with basic elements).

The example outlined in Fig. 15.22 is a mechanical process of two types of shafts in a pulsed production system. Two types of shafts are provided, with their own amounts and average arrival times. They are combined in a common pulsed material flow for milling and lathing. The material flow is split regarding the

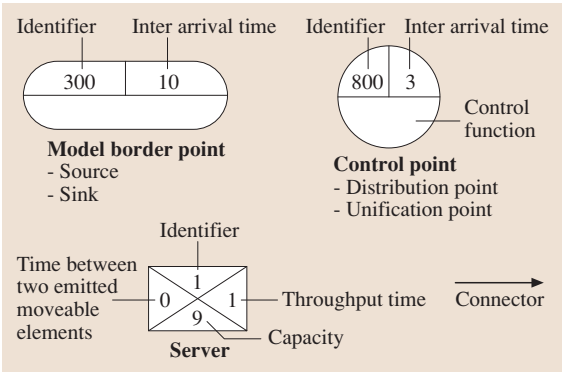


Fig. 15.21 Basic modelling elements for a simulator

shaft types because each type of shaft has its own finishing process. After finishing, the material flows are combined and pulsed again for final washing and sales.

The process will be simulated by a discrete/event-oriented simulation system. As well as the structure, data for system control may also be provided.

EPC. The event-driven process chains (EPC) is a general method for business process modeling. The essential constructs are: function, event, connector, and (sub)process chains (Fig. 15.23).

Besides the presentation of the control flow, the analysis of data flow is the subject of interest. This can be done via the input/output assignment of the information objects in a data flow model. An extension of the event-controlled process chains with the aspect of organizational responsibilities can be performed through the association of organization units to functions. Due

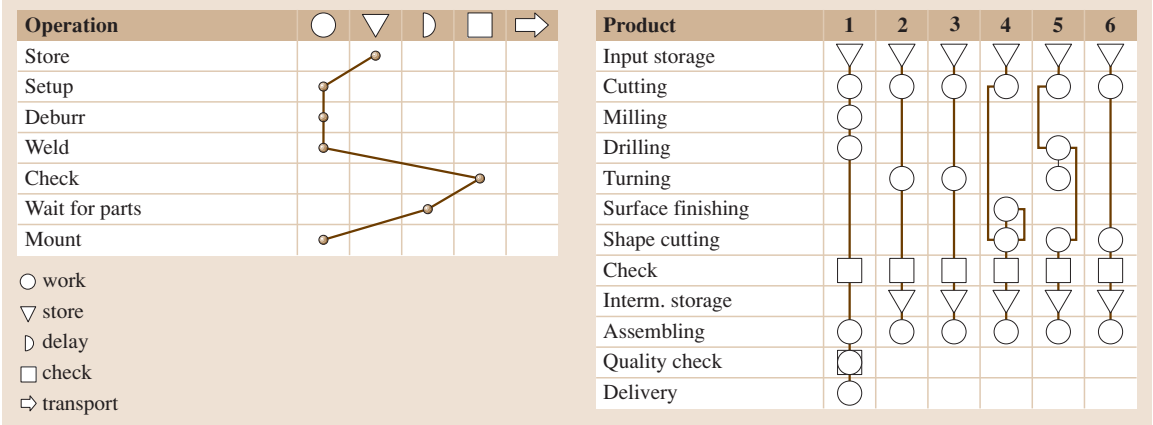


Fig. 15.20 Terminology and symbols for operations modeling

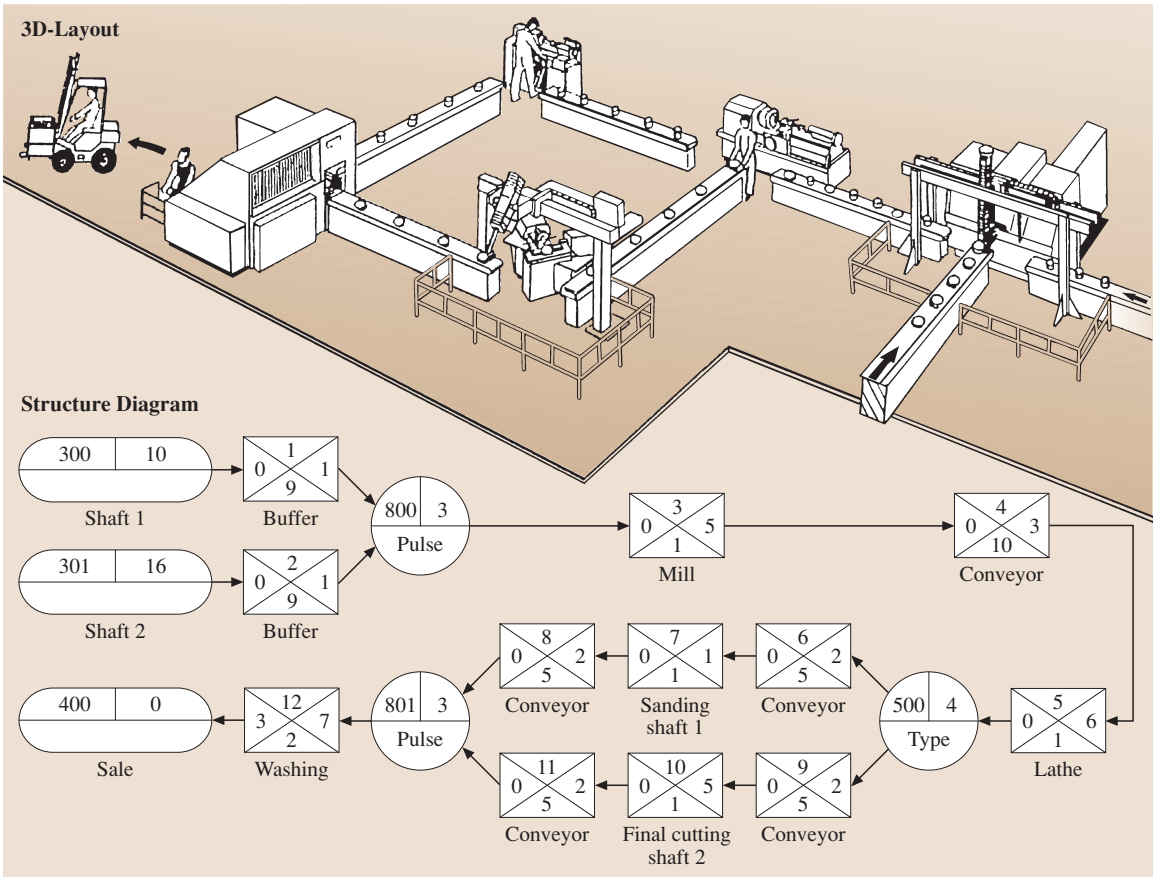


Fig. 15.22 Layout and structure diagram of a shaft manufacturing (after [15.62])

to the lack of an explicit presentation of process instances, the simulation of modeled business processes is difficult [15.53]. EPCs are explicitly designed for the modeling of business processes. Therefore they have individual elements of presentation of all related aspects. Thus, business models presented with EPCs are very demonstrative (Fig. 15.24).

Precise standards for business modeling with EPC are available.

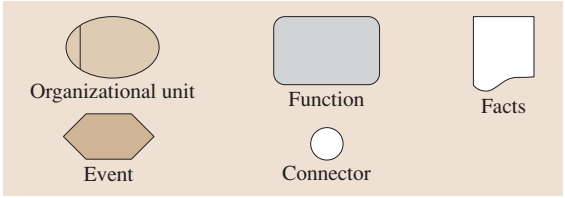


Fig. 15.23 Constructs of event driven process chains

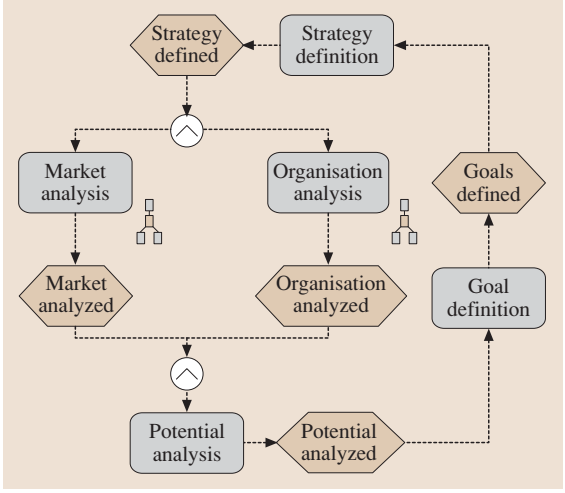


Fig. 15.24 Strategy finding process in EPC

ERM. The most widespread concept method for semantic data models is the entity relationship model (ERM) by *Chen* [15.63], which presents the relations between data units (entities) graphically. The entity relationship model is based on a classification of the real world into entities and relations. Thus, using an ERM, information flows can be modeled, setting events and processes (functions) in relation to each other. States become visible due to the inclusion of information technology resources into the model. Partly, organization units can be defined.

The main information in an ERM is provided by the presentation of relations and their occurrence between the objects. All forms of relations (1:1, 1:*n*, or *n*:*m*) can be presented. To keep practical applications simple, it is useful to insert artificial objects and to convert real relations into artificial, idealized system elements for the simplification of relationships. Multiple 1:1 or 1:*n* relations are usually easier to handle than one *n*:*m* relation.

Entities are objects or issues that can be immediately identified; relations are associations between entities. For example the project work can be a relation between two entities of type person and project (Fig. 15.20).

Data input is done using instances of entities, i. e., a certain characteristic of entities or relations is specified by a set of attribute value couples; for example, if *name* is an attribute of the entity *employee*, *Peter* is a value of this attribute in an instance of the entity. In an extended approach, information can be assigned to relations using attributes. The entities of a model are classified into different entity sets. The information about entity instances of an entity set are gathered in tables called an entity relation. Every row of the table is called an entity tuple. In entity relationship diagrams every entity is presented by a box; every relation is presented by a rhomb. The role of entities in the relation is static (Fig. 15.25). The method of ERM is used to illustrate complex contexts and if a reduction of complexity is intended. The ERM produces uniqueness.

a)

Attribute	Employee-nr.	Name		Age
Domain	Employee-nr.	Given name	Surname	Years
	1522	Peter	Meier	30
	1590	Fritz	Schultz	53
Entity-tuple	2007	Petra	Schwartz	22
	.	.	.	.
	.	.	.	.
	.	.	.	.

b)

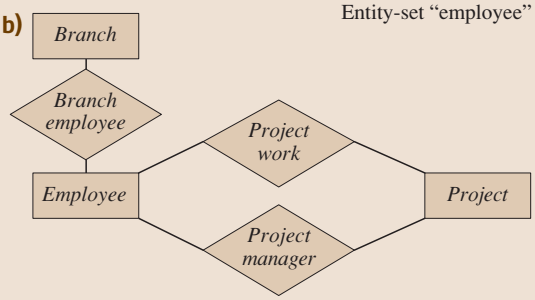


Fig. 15.25a,b Example of an entity set end an entity-relationship-diagram

Modelling Techniques

Modelling techniques provide either a syntax of the model entities or a semantics defining the modelling process.

With UML and SYSML two syntax defining models and with RUP a semantic based modelling process relevant within organizations modelling are considered.

UML. The unified modelling language is a language for specification, visualization, construction, and documentation of models for software systems and business models (in its specific sense) as well as systems (in its general sense). It is based on the unification and improvement of object-modelling technique (OMT, *Rumbaugh*), object-orientes software engineering (OOSE, *Jacobson*), and the method of *Booch* [15.64].

UML is a modelling language and notation but intentionally not a method. UML can be the base for various methods because it provides a defined set of modelling constructs with unique notation and syntax.

The current version 2.0 of UML defines a wide range of diagrams for modelling. They can be distinguished as structure models and behaviour models. The most important are (see also Fig. 15.20):

- The use case diagram applied to show users, use cases, and their relations and to describe the outer system behaviour in its recognition by a user.
- The class diagram shows classes with its encapsulated properties and data (attributes), its encapsulated behave-our (operation/functions), the relations between classes, and the structural dependencies among classes.
- The activity diagram shows activities, object states, states, state transitions, and events and, thereby, describes object behaviour using their activities, which are performed during execution.

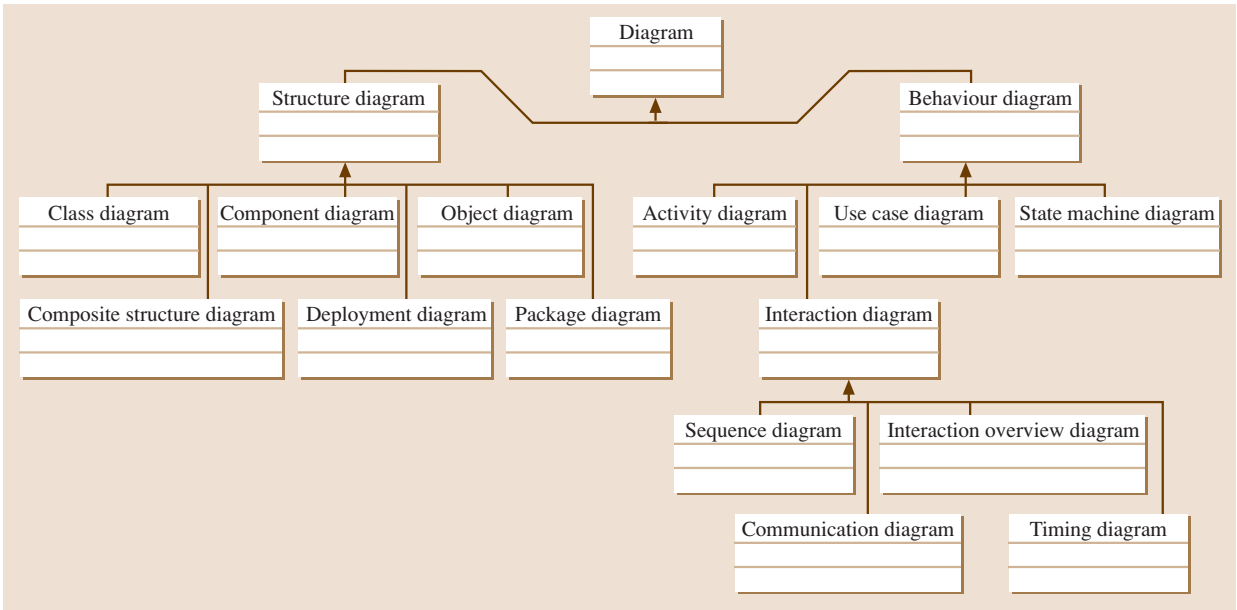


Fig. 15.26 UML diagrams

- The collaboration diagram shows objects and their relations including their exchanged messages and, thereby, describes the acting model objects for a certain delimited context.
- The sequence diagram shows objects and their relations including their exchanged messages and, thereby, describes the chronological sequence of interactions between a set of objects.
- The state diagram shows states, state transitions, and events of objects and, thereby, describes the distinct states of an object as well as the functions which cause the state transitions of this object.
- The timing diagram shows states, state transitions, and events of objects in its temporal dependencies and, thereby, describes the temporal behaviour of objects.
- The component diagram describes system components consisting of different model objects and their relations.

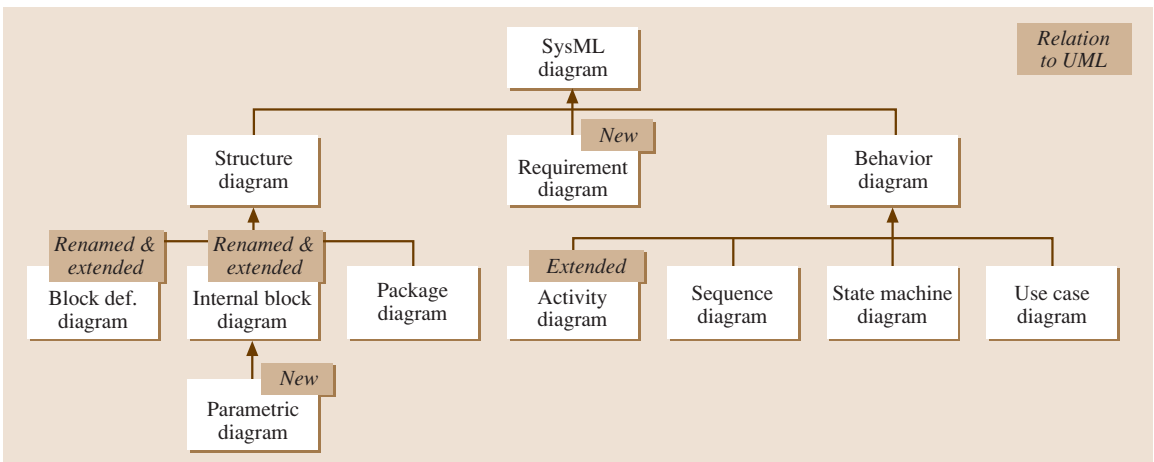


Fig. 15.27 SysML diagrams

- The deployment diagram shows components, nodes (devices/resources), and their relations and describes, which system components run on which devices.

UML enables the modelling of various systems from various points of view dependent on the intentions of the modeller. It is not related to a certain field of application.

SysML. The systems modelling language (SysML), is a modelling language targeted for the systems engineering domain. It is a visual modelling language supporting the specification, analysis, design, verification and validation of a broad range of systems and systems-of-systems.

SysML has been initially developed as an open source specification project. Now SysML has been adopted by OMG as international standard [15.66].

The foundation of SysML is a subset of UML which has been extended by systems engineering relevant modelling capabilities as depicted in Fig. 15.27.

The diagrams which are applied unchanged from UML are package diagram, sequence diagram, state machine diagram and use case diagram.

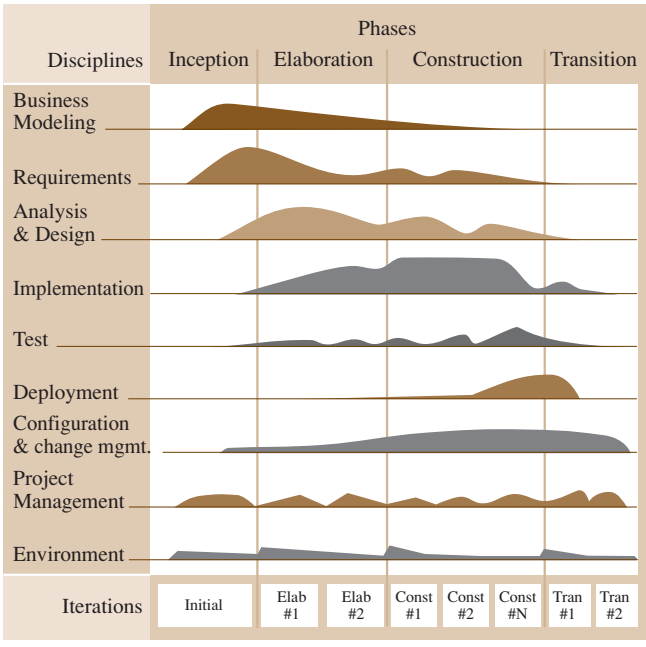


Fig. 15.28 Phases of the rational unified process in relation to hte activities made within them [15.65]

The renamed and extended diagrams are the block definition diagrams and the activity diagram.

Block diagrams are based on UML class and UML composite structure diagrams. They are based on the term of blocks which are the basic structural elements in SysML. Blocks describe the structure of elements or systems based on multiple compartments providing the block characteristics like properties, operations, constrains, allocations to the block, and requirements the block satisfies. The block definition diagram (bdd) describes the relation among blocks while the internal block diagram (ibd) describes the internal structure of a block.

SysML extensions to activities are the support for continuous flow modeling and its alignment of with enhanced functional flow block diagram (EFFBD).

Within both types of diagrams, block as well as activity diagrams, a (with respect to UML) new modeling concept is integrated representing communication among entities (blocks or activities). SysML specifies interactions points on blocks and parts called Ports. SysML has two port types, standard ports and flow port.

Facing the needs of systems engineering within SysML system requirements can be modelled and provided within the requirement diagram. Thereby, system analysis and evaluation can also cover requirement problem solution as well as description and evaluation of alternatives.

The second extension is the parametric diagram usable for the specification of the dependency of parameters within different entities of a system.

Similar to UML SysML is not related to a specific application domain. It provides a modelling syntax applicable within various fields.

RUP. The rational unified process (RUP) is initially an iterative software development process [15.67]. It is not a single fixed process, but an adaptable process framework. It can be tailored to the development of organizations and software design projects by selecting elements of the process that are appropriate.

The unified process was designed from the start to include both a generic, public domain process (known as the unified process), and a more detailed specification known as the rational unified process which could be marketed as a commercial product.

The RUP is based on a lifecycle consideration of a system design and implementation process within a spiral model. It has been created by assembling the content elements into semi-ordered sequences. Hence,

the RUP lifecycle is available as a work breakdown structure, which has to be customized to address the specific needs of a project. The RUP lifecycle assumes that a project has four phases (Fig. 15.28):

- Inception phase: In this phase the business case which includes business context, success factors, and financial forecast is established and complemented by basic business use case models, project plans, initial risk assessments, cost assessments, stakeholder considerations, project descriptions, and requirement descriptions.
- Elaboration phase: in this phase the problem domain analysis is made and the basic architecture of the intended system is developed. This is done by developing use-case models in which the use-cases and the actors have been identified, identifying significant use cases, and revising business cases and risks.
- Construction phase: In this phase, the main components and features of the intended system architecture and behaviour are designed. This phase is the main design phase developing the concrete system implementation.
- Transition phase: In the transition phase, the product will be moved from the development organization to the end user, i. e. the product will be delivered. In the case of an organization or a complex system the organization/system will be established.

The RUP provides a methodology for organization/system design fixing semantics of the results of different design phases. Intentionally it is not connected to special models and modelling techniques. Nevertheless, very often UML and SysML (depending of the intended domain of the result) are applied within RUP execution.

15.5 Enterprise Resource Planning (ERP)

Enterprise resource planning defines the task of enterprises to optimise plans for the most efficient input of all available resources. As these activities are generally computer supported, the term of ERP is actually used to describe software or software implementations to operate and optimize a number business processes in companies [15.68].

Enterprise resource planning systems attempt to integrate all data and processes of an enterprise organization into a unified software system. Formerly used terms have been also production planning and control (PPC), materials requirement planning (MRP) and manufacturing resources planning (MRP II) [15.69].

Revolving plans are generated for the production programs as well as for the allocation and expected demand of resources like production capacities or inventory. Forecasts and customer orders (primary demands) are broken down to the finished goods', aggregates' and parts' levels (secondary requirements) using a set of well established procedures (net requirements calculation, lot sizing calculation). Moreover ERP also includes a number of administrative functions, as accounting or billing.

In order to give an overview on ERP functions, the key procedures that are implemented shall be detailed in formal terms that clearly differentiate the resources regarded.

15.5.1 Resources and Processes

As all software systems, also enterprise resource planning systems are based on models of the business processes regarded. The easiest way to built up these models is to use the given information base in the company. As all companies use bills of material (BOMs) and work sheets or bills of operations (BOOs) the information concerning:

- Machine (cut, drill, mill, deform etc.)
- Move (transport, insert, eject etc.)
- Unite (assemble, weld etc.)
- Separate (saw, split, etc.)

may be read out and used for planning and process modelling.

Figure 15.29 shows a manufacturing flow model on the basis of BOM and BOO information. In ERP, a number of worksheets and bills of material are interlinked in order to obtain routing, process, and flow models as the data input for enterprise resource planning and control [15.70, 71].

In order to define the full manufacturing processes all resources are assigned to functions and operations that are necessary. In the ERP context, resources are workforces, skills, objects, and equipment. ERP uses

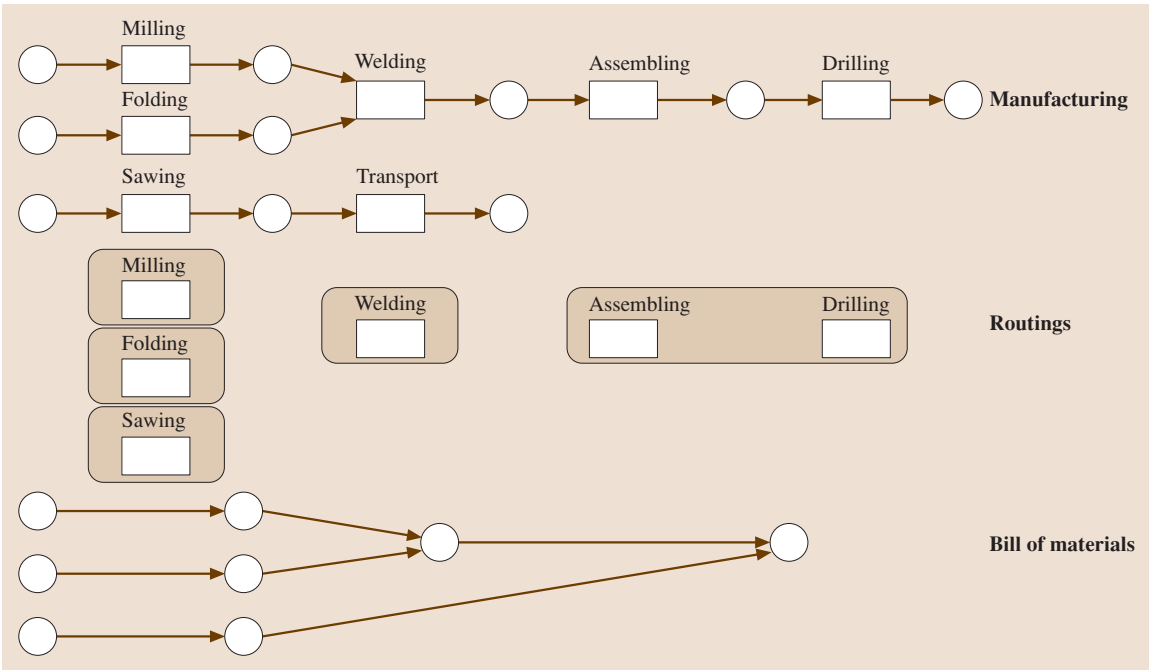


Fig. 15.29 List of date and models used by ERP

a basic set of resources R to model the manufacturing processes, where R may be split into subsets

$R = R_1, R_2, R_3, R_4$ meaning:

R_1–materials, parts, and final products (objects),

R_2–machines, lines, (and equipment),

R_3–workforce (human resources),

R_4–auxiliary equipment, tools, and measuring technology.

For every unit of these resources, inventories and availabilities may be forecast, planned, and observed over the time horizon.

15.5.2 Functionalities of ERP Systems

Enterprise resource planning (ERP) software solutions [15.72] are designed for systematic disposition and planning of all enterprise resources in production, distribution, logistics, finance, and personal. Different kinds of ERP systems offer varying degrees of functionality and the properties depend on the areas focused on, for instance, manufacturing business or service business [15.73]. Their primary objective is the optimization of business processes in product development, materials logistics, production, maintenance, quality assurance and service, sales, and distribution (Fig. 15.30). Key ingredients of ERP systems are unified databases to store the data for the various system modules.

Thus ERP systems are universal standard software solutions to control and optimize business processes. Whereas early systems focused on intra-enterprise processes, interenterprise processes are also covered, and ERP systems are becoming web enabled with additional software modules, e.g., for supply chain management (SCM) and customer relationship management (CRM).

In manufacturing companies, the core of ERP is provided by software functionalities for production planning and control. This includes all customer demands indicating production areas as well as the demands resulting for subsequent planning steps, demand forecasts for products, and derived demands for aggregates or parts [15.74]. Deeper studies of ERP, especially current research projects in the field [15.75], focus on the data structures and the logical steps behind ERP procedures.

15.5.3 ERP Procedures

Enterprise resource planning systems are implemented to model, plan, and control resources that are required

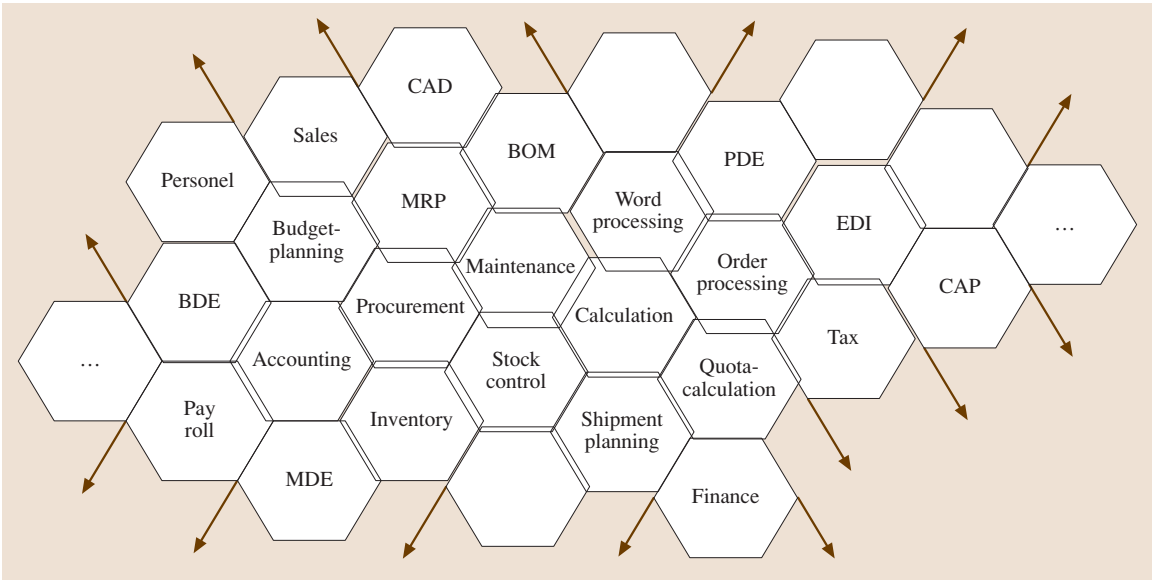


Fig. 15.30 Functionality configuration of an ERP-system (CAP – computer-aided planning; EDI – electronic data interchange) (after [15.76])

for the output and throughput of a company. The availability of resources is forecast and monitored over discrete-time axes (calendar dates, shift numbers, or hours).

Discussion about the data used and the logic steps of ERP may be based on the term

$$\text{ERP} := [R, S, T, (Rb(t), r(t), D, E, A)] ,$$

where R is a set of resources, A is the procedure pointer, S represents the relations between resources, E is a decision, T is the time scale, D is the diagnosis, $b(t)$ is the requirement time function per resource, and $r(t)$ is the availability time function per resource.

These elements constitute the pillars of all ERP core procedures according to Fig. 15.31. Forecasting is done along discrete time units (time scales are weeks, days, and shifts), delivering revolving plans for demands, order loads, inventory levels, and capacity availabilities. These plans are generally output as data columns and histograms.

Statistical Inventory Control (SIC)

The SIC procedure determines forecasts for inventories and orders by extrapolating past product demands and resource consumption (on a stochastic basis). Most frequently the methods used for forecasting are arithmetic averaging (in some cases weighed averages) and exponential smoothing. Based on actual and planned levels of inventories, demands are summed and transformed into orders. Volatilities in demand and errors in the forecasts are buffered by safety stocks; recommended levels are determined by statistical methods based on the variance parameter of the normal distribution of fulfilled orders.

Statistical inventory control (SIC) is applied to purchase materials and low-price parts as well as for internal company multistage requirement planning, in some cases to ensure the availabilities of equipment

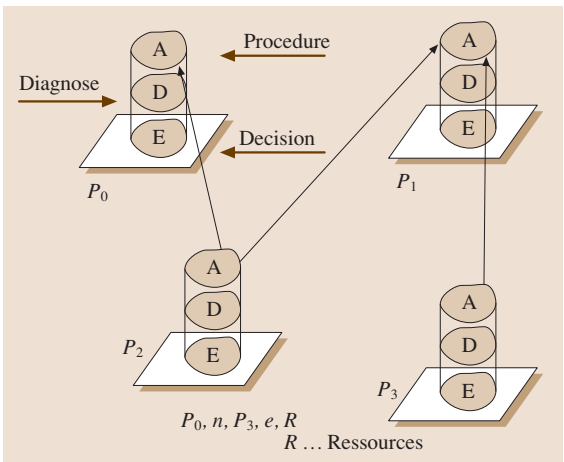


Fig. 15.31 ERP-model-basics

Table 15.9 Description of important ERP procedures (SCH – scheduling; EOQ – economic order quantity)

ERP procedures	*R*	*S*	*T*	*b*(*t*)	*r*(*t*)	Operators *D*	*E*	*A*
SIC	R_1	–	Discrete	Yes	Yes	Material requirement? Inventory?	Order EOQ	Next low inventory level
MRP	R_1	BOM	Discrete	Yes	Yes	Material requirement?	Order EOQ	Next BOM position
MRP II	R_1, R_2, R_3, R_4	BOM	Discrete	Yes	Yes	Resource requirement? Resource inventory?	Order release	Resource list
SCH	R_1, R_3, R_4	Time dependencies	Discrete	–	–	Order due date feasible?	Order release order split/ concurrent procedure	Subsequent time span (forward/ backward)
CSP	R_1, R_2, R_4	Working plan	Discrete	Yes	Yes	Resource requirement? (time, capacity, lot)	Load shift sequencing queuing	Subsequent process step
Kanban	R_1	–	–		–	Minimum inventory?	Card release	Next card

(e.g., tools). SIC application does not assume a complete bills of materials or other information detailing the setup of products or aggregates.

Material Requirement Planning (MRP)

The MRP procedure may consider all materials and parts that are listed in the bill of materials (BOM); the interrelations set up between the products are used for the computation of the requirements. A (given or forecasted) production programme is broken down into requirements at all levels of the product (on a deterministic basis). In order to determine throughput times and accurate quantities, delays and waste factors are added to the BOM relations (calendar dates, delay times, lot sizes, volumes, and economic order quantities). Generally the products, parts, and aggregates (BOM positions) that are considered by MRP are expensive, difficult to supply, or discontinuous in demand.

Two alternative modes are applied:

- Order-driven MRP, which breaks down the demand per order
- Program-driven MRP, which breaks down the demands for all BOM positions, using complex net requirement calculations on the basis of inventory levels

Manufacturing Resource Planning (MRP II)

As sketched above, MRP II extends MRP to interenterprise contexts. It uses universal descriptions independent of industry branches [15.77]. Planning and control is applied to multistage value chains. MRP II is characterized by a strong material focus as well, strictly separating lot sizing and capacity loading [15.78]. All process steps are planned and scheduled. Capacity planning activities are a part of standard software activity.

Scheduling Procedure

For an approximate determination of due dates and dates for execution of process steps the time consumptions to be expected for the order executions is considered. Resulting schedules for R_1 (materials, parts, final products, etc.), R_2 (machines, lines, etc.), and R_4 (auxiliary equipment, tools, measuring technology) anticipate order allocations as well as order execution and finish constellations (Gantt charts) [15.79]. One-of-a-kind production may require more elaborate (graph-theory-supported) procedures such as the critical-path method (CPM), the project evaluation and review technique (PERT) or the metra potential method (MPM) [15.14] for scheduling. The benefits of production scheduling include setup cost and inventory reduction, increased production efficiency, labor load leveling, and real-time order information.

Capacitated Scheduling (CSP)

Capacity requirement and scheduling generates detailed plans and schedules for machines, equipment, and materials, feasible to be executed in manufacturing areas [15.80]. A special case is the capacitated lot-sizing lead-time problem (CSLP), where smooth loads are generated by time and/or assignment shifts using sequencing algorithms as well as order-split logics and queuing rules.

Kanban

Kanban is a procedure that controls the flow of material by *pulling* parts as needed through the manufacturing system. A part is manufactured if and only if a part of this kind has been taken out of the buffer. The procedure allows minimum inventory, ensuring highest order availability at the same time. Kanban (Japanese for "card") is used for the control of mass production and large series throughputs. For each part a standard lot size and minimum inventory level are defined, represented by the number of cards released into the process.

Using the ERP formalization as introduced above, all described procedures may be categorized as shown in Table 15.9.

15.5.4 Conclusions and Outlook

ERP applications are widespread in enterprise organizations. Applied properly, they provide many advantages. Nevertheless the most elaborate ERP procedures also exhibit weaknesses regarding flexibility (dynamic routing or rush orders) and interoperability (with other systems, such as MES). Problems are caused by newly arriving high-priority production orders, since the running production cannot be modified. ERP implementations are often seen as too rigid and too difficult to adapt to the specific workflow and business processes of some companies [15.81].

Also, in areas related to software implementation, expected ERP benefits may not be fully attained as:

- Customization of the ERP software is limited. Some customization may involve changing of the ERP software structure, which is usually not allowed.
- Re-engineering of business processes to fit the *industry standard* prescribed by the ERP system may lead to a loss of competitive advantage.
- Many of the integrated links need high accuracy in other applications to work effectively. A company may achieve minimum standards, then, over time *dirty data* will reduce the reliability of some applications.
- The system may be overengineered with respect to the actual needs of the customer.

To improve ERP applications, potential is seen in more intensive cooperation [15.82]. Concepts such as built-to-order supply chain (BOSC), efficient customer response (ECR) [15.83], and continuous replenishment planning (CRP) [15.75] are discussed, aiming for smoother flows. Multiple forecasts within supply networks, changing demand patterns, and general synchronization problems call for collaborative planning, forecasting, and replenishment (CPFR) [15.84]. Nevertheless these approaches, even if accompanied by firm alignments via very rigid standards, as proposed by the Voluntary Interindustry Commerce Standard Association (VICS) [15.83] are still considered immature [15.85]. Their operation is laborious, as they ignore fundamental problems of collaboration involving diverse self-interested actors with conflicting motivations.

With changing organizations in enterprises, ERP logics and software solutions have to be redeveloped too. Currently the vendors of ERP software are working on solutions to avoid the shortfalls listed above [15.85]. The developments are oriented towards advanced industrial organization concepts [15.86], greater flexibility, and variability of product and process data [15.87]. First solutions draw from networked control concepts and the latest ICT devices, e.g., radiofrequency identification (RFID) and multi-agent systems (MAS).

15.6 Manufacturing Execution Systems (MES)

To manage the optimization and control problems accruing in larger manufacturing systems, which also include large amounts of numerically controlled (NC) equipment, manufacturing execution systems (MES) have been designed. Presently, manufacturing execution systems incorporate algorithms for short-term production planning and control. According to the specification of the Manufacturing Enterprise Solutions Association (MESA) [15.88], MES covers 11 functionalities, as listed in Table 15.10.

For efficient implementation of these 11 functions several specific architectures have been developed.

Table 15.10 Main functions of MES

MES functionality	Attributes
Resource allocation and status	• Shows resources, including reservation and dispatching of machines, tools, labor skills, materials, and other equipment • Provides read/write access to documents and data that must be available to enable the necessary work
Operations/detailed scheduling	• Performs sequencing of resource operations and functions based on priorities, attributes, characteristics, and/or recipes • Reflects specific features of the scheduled resources such as necessary activities for process setup and possible process sequences
Dispatching production units	• Prepares the flow of units to be produced such as jobs, orders, batches, lots, and work orders • Creates dispatch information presenting the sequence
Document control	• Manages the control of information containing units (forms, files, records, etc.) relevant for a manufacturing unit
Data collection and acquisition	• Provides an interface to collect all data such as interoperational production data and parametric data that are included in the controlled documents of a resource
Labor management	• Manages status information of personnel, including time and attendance reporting as well as certification tracking
Quality management	• Provides real-time analysis of measurements collected from the resources before, during, and after the manufacturing process
Process management	• Monitors the manufacturing process resources and automatically corrects resource behavior or provides decision support to operators
Maintenance management	• Controls activities to maintain resources to ensure their availability for manufacturing and ensure scheduling for periodic or preventive maintenance
Product tracking and genealogy	• Provides visibility of the current status of the manufacturing processes and the current state of resource activities • Creates a historical record
Performance analysis	• Provides up-to-the-minute reporting of actual manufacturing results and its comparison with past history and expected business results

The most popular class of systems focuses on resource allocation, scheduling, and dispatching on the one hand, and data collection and acquisition on the other hand. As the overall complexity of MES is increasing (as manufacturing systems become more complex), architectures have been adapted with distributed functionalities and to develop distributed MES systems [15.89].

Special effort has been invested in the field of intelligent distribution of the necessary decisions. One major technology used within this field is agent-based systems. All of these efforts reflect similar structures dealing with negotiating orders and resource entities [15.90].

Another field of major interest is the interfacing of data sources relevant for MES system behavior. An advanced architecture in this field is the open robot interface for the network (ORiN) – one of the most important data collection and acquisition architectures – which offers interoperability of different functionalities with the possibility of integrating systems from different vendors.

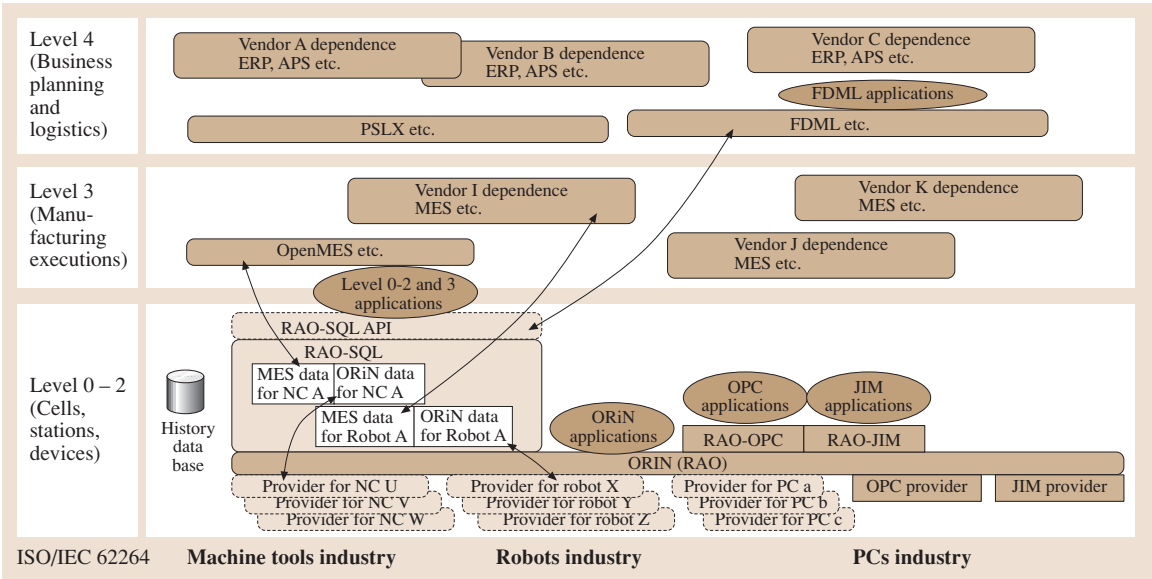


Fig. 15.32 Concept of the information-interoperable environment (FDML – flow description markup language; OPC – OLE for process control; OLE – object linking and embedding; PSLX – planning and scheduling language on XML specifications; RRD – robot resource definition; RAO – robot access object; SQL – structured query language)

15.6.1 Information-Interoperable Environment (IIE)

The open robot interface for the network (ORiN) [15.91] is an example of an open technology, developed by the ORiN consortium for robots. It has been extended through the development of gateway systems for other standards such as open connectivity via open standards (OPC) [15.92] and is currently active in the ORiN consortium. Therefore, the ORiN technology has been applied not only to robots and programmable logic controllers (PLCs) but also to machine tools. Additionally, an information-interoperable environment (IIE) that can interlink device-level information and MES-level information has been developed using ORiN.

A mechanism that can acquire device-level information and MES-level information corresponding to the needs of the user for increasing productivity on a daily basis is needed. Such an IIE has been developed using ORiN's middleware solution RaoSQL (RAO – robot access object; SQL – structured query language) [15.93], as shown conceptually in Fig. 15.32.

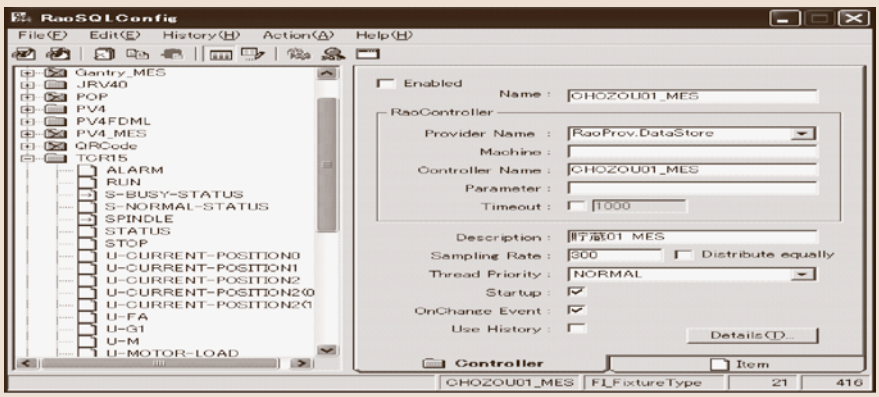


Fig. 15.33 Configuration tool of the environment

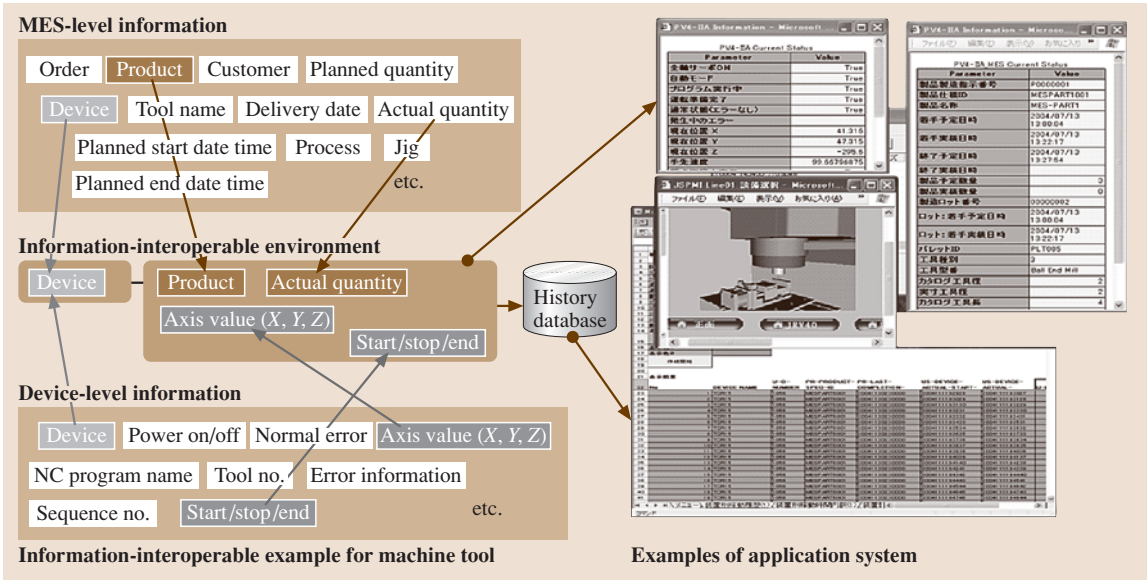


Fig. 15.34 Concept of applications

The RaoSQL has a RaoSQLController class for discriminating amongst different devices and a RaoSQLItem class for defining various types of information from the devices. Information corresponding to the robot access object (RAO), which is an ORiN model, can be defined in RaoSQLItem objects. That information can then be provided to application systems through the RaoSQL application programming interface (API). Additionally, the data variance history can be logged to market databases. In addition, RaoSQLItem

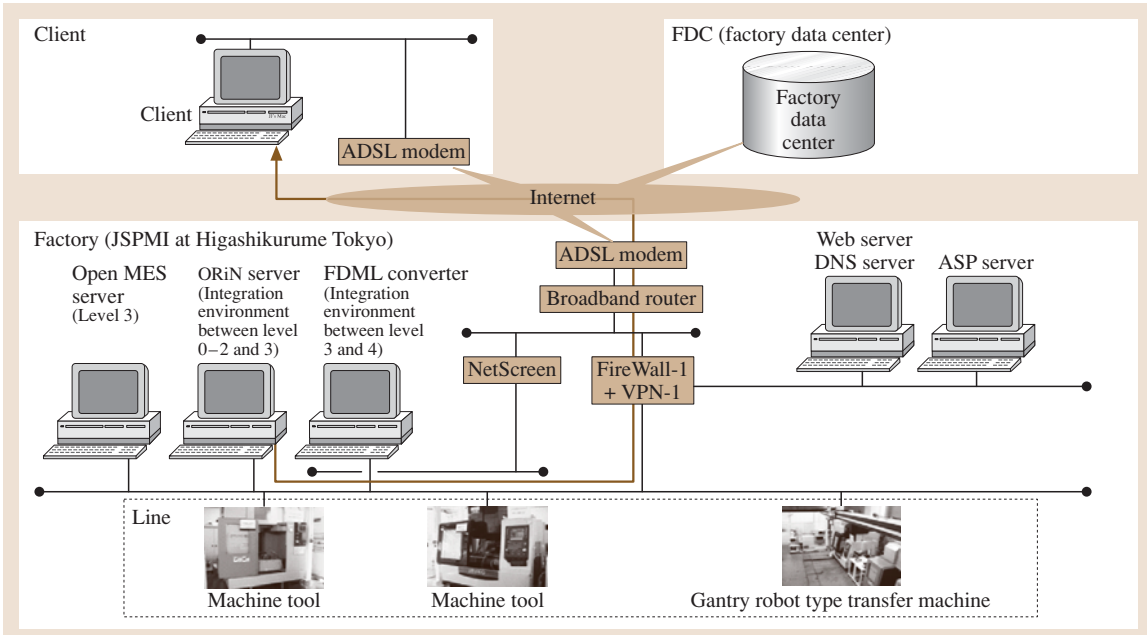


Fig. 15.35 System configuration of the prototype

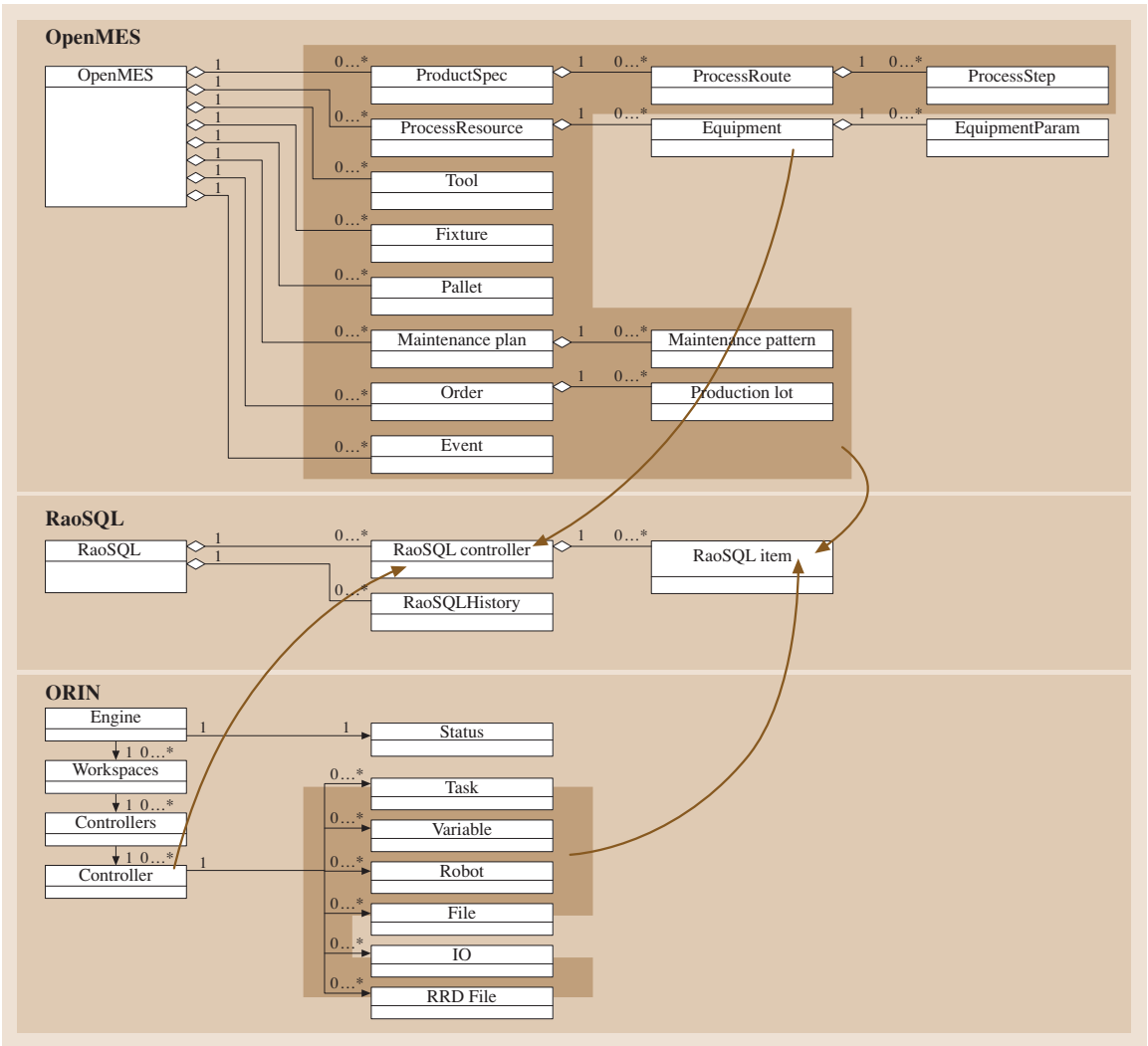


Fig. 15.36 Integration mechanism between Open MES and ORiN

objects can provide not only device-level RAO information but also MES-level information using the data store function of RaoSQL that handles optional information. Information exchange with the MES level of this IIE uses the RaoSQL API. Moreover, the configuration management tool of the IIE can manage both the RaoSQLController and RaoSQLItem objects. Dynamic reconfiguration by the user in order to make productivity enhancements is possible through the use of the configuration tool, as shown in Fig. 15.33. An application illustration using this system is shown in Fig. 15.34.

By using the IIE with a web server, remote monitoring of device- and MES-level information through web applications can be easily customized by the user. With the accumulation of data available from the device- and MES-level database histories other application systems such as an operation management system can be created using Excel, giving users even greater freedom in the computing environment. Custom-made applications for the purpose of increasing productivity are available through these methods without having to resort to commercial software.

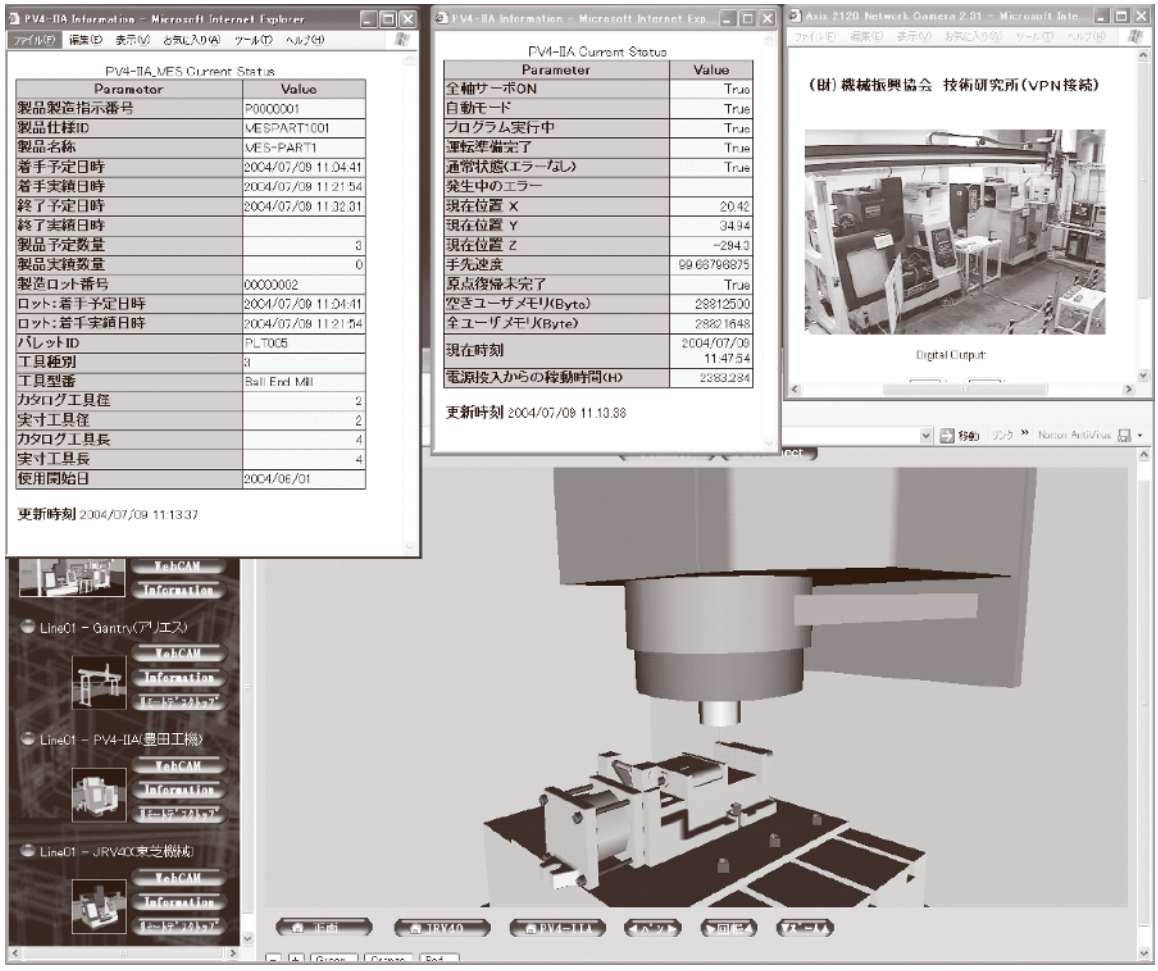


Fig. 15.37 Test operation of the environment

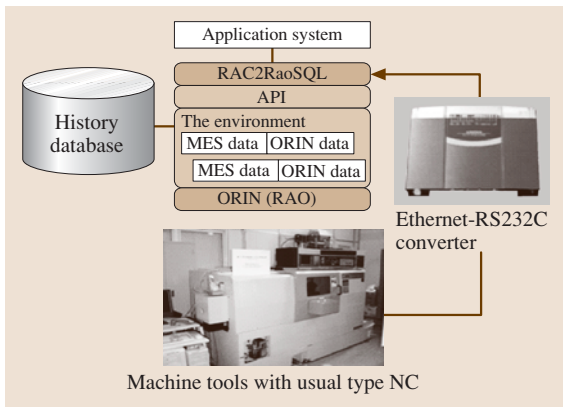


Fig. 15.38 Connection method used for usual type NC

Fig. 15.39 An example of application system

15.6.2 Development of Prototype Application Systems

An Application System for Open NC Machine Tools and MES Applications

For a manufacturing system line consisting of one transfer robot and two machine tools, integrated with OpenMES [15.94] as shown in Fig. 15.35, the mechanism of the interoperable system is shown in Fig. 15.36. The equipment class of the OpenMES has been mapped to the RaoSQLController class of RaoSQL. The other OpenMES classes and other ORiN classes have been mapped to the RaoSQLItem class of RaoSQL.

Provider software as interface software between devices and ORiN for the machining tool and the transfer robot have been developed using the APIs of each controller with Microsoft's Visual C++. However, because OpenMES uses Java, a software interface between OpenMES and RaoSQL using a Java wrapper was developed.

Test operation of the prototype was performed using the three-dimensional (3-D) layered information viewing environment (3-D LIVE) [15.95]) which was developed in advance. Conventional 3-D LIVE was developed as an application system that uses RAO information directly. The ability to show MES-level information together with device-level information became possible by changing the information reference source in 3-D LIVE to the IIE, as shown in Fig. 15.37.

Application System for Machine Tools Without Network Interface

Examples of accessing operational information from machine tools without network interfaces without a major conversion process include:

- Acquisition of operational status through the machine tools operation panel/switchboard or digital input output (DIO)
- Estimation of a machine tool's cutting load by measuring the electrical current of the spindle amplifier

NC controller comments as well as variables in the external output command of the NC program can be exported through the RS232C interface of the NC controller through the use of macro functions. Additionally an external direct numerical control (DNC) computer can manage the operational status of a machine tool if that machine tool is running a DNC system.

Accessing information using an external output with macro functions may be adopted for machine tools without a network interface. Ethernet-to-RS232C converters can be used to integrate this system as a very economical solution. The corresponding system configuration is shown in Fig. 15.38. There are also cases (such as when using a CAD/CAM system) where MES-level information such as product name, date of delivery, and tool information are known right after the NC program is made. In such cases, the external output commands from the macro function for the MES-level information can be inserted directly into the NC program. The format of the comment information from the external output is based on the robot action command (RAC) [15.96] specification that was developed by the ORiN consortium. Once the NC program has been prepared it is transmitted to the NC controller of the machine tool. The comment information can then be used for setting up the machine tool by viewing the display panel of the NC controller. The comment information is then output through the RS232C interface of the NC controller, which is synchronized with the execution of the NC program. Subsequently, the comment information is sent to the RAC2RaoSQL, which is an application system for receiving the RAC data through a transmission control protocol (TCP)/internet protocol (IP) socket or virtual COM port on a network using the Ethernet-to-RS232C converter. After the RAC2RaoSQL receives the comment information it is written to the IIE.

The configuration of the test system consisted of a factory site and a remote site. The factory site had a developed prototype system for a turning machine with an IIE. A client computer, using a web browser, was prepared at the remote site and connected to the factory over the internet using a virtual private network (VPN). Figure 15.39 shows an example of the indicator screen on the client computer. The NC program's name, executing sequence number, the product's name, date of delivery, scheduled production count, actual production count, and the process name of the executing NC program can all be seen in this example.

15.7 Advanced Organization Concepts

Continued trade liberalization and reductions in trade barriers are accelerating international flows of information, goods, and services. The major impact on enterprises has been the greater availability of resources as low-cost labor and manufacturing capacity, which has become an increasingly compelling reason to move towards sourcing parts and components globally. Other key driving forces can be identified as shortening product lifecycles, placing a premium on speed to market, and profiting from rapid declining costs of transportation and communication.

Companies, being cornered by these phenomena, have realized that just tuning operations and functions will not remove the difficulties. More integrated organization concepts, supporting the improvement of key performance indicators such as business process reengineering (BPR), the improvement of quality by total quality management (TQM), the improvement of work in process by just-in-time (JiT) production, and the improvement of productivity by total productive maintenance (TPM) have been regarded as sustainable solutions. However, these concepts only represent important steps towards rediscovering the organizations' strategic nature in enterprises. In order to face new global challenges, adequate organization structures turn out to become the most powerful and indispensable strategic instruments for companies (Fig. 15.40).

Fierce competition at all levels, first observed in the *triad* of North America, Europe, and the Far East and later experienced in many parts of the globe, enforces holistic thinking in industrial organization. Expanded objectives and volatile objective settings demand innovative, learning, open, networked, flexible structures. The assumption that industrial organizations are results of predictability and deterministic design no longer holds. Hierarchical organization structures become obstacles to market success.

More theoretically speaking, organizations and operations that are exposed to such complexity [15.98] induced a paradigm change [15.99], as their behavior and reaction become nonpredictable and do not coincide with proven and known patterns of management. Organizations and structures have to be redesigned and redirected; successful and sustainable solutions draw from generic approaches for new enterprise organization.

Lean manufacturing (lean production) is one such approach; later others appeared: the fractal company, agile manufacturing, and holonic and bionic manufacturing.

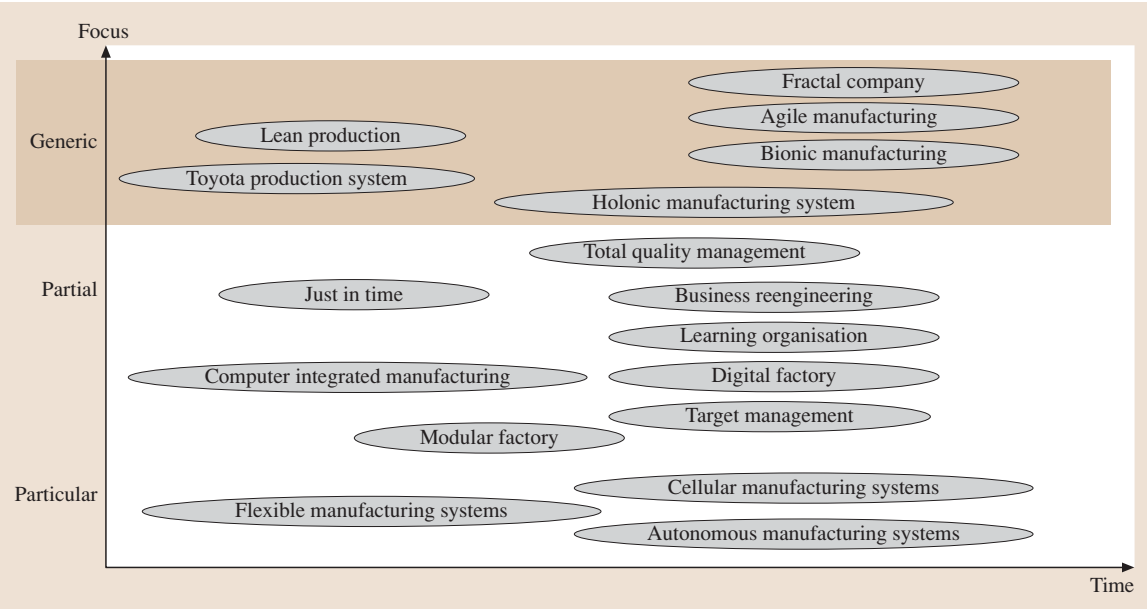


Fig. 15.40 Important management approaches due to paradigmatic shift (according to [15.97])

15.7.1 Lean Production

Since the MIT studies on *the machine that changed the world* [15.100] lean production has become one of the most frequently discussed organization concepts.

Lean production is an assembly-line manufacturing methodology developed originally for the manufacturing of automobiles, derived from the theory of constraints [15.101]. It is also known as the Toyota production system (TPS) [15.102]. Many practical approaches that are also referred to as lean manufacturing (lean management, lean company, lean enterprise etc.) are motivated by this methodology, or originate from Toyota production system blueprints. The guideline for lean production can be formulated as *to get the right things to the right place at the right time, first time, while minimizing waste and being open to change*. Lean manufacturing can be classified as a team-centered organization specification. While developing the principles of lean production [15.103], it was discovered that, in addition to eliminating waste, the methodology leads to improved product flow and better quality.

There is no established theory for lean production, but theoretical grounds may be found in the theory of constraints. However detailed methods and rules for the successful operation of *lean* are outlined. The most cited rules (simplified) are:

1. Produce only what is ordered and only when it is ordered. Apply this to product, organization, as well as product features; otherwise there will be waste.
2. Investigate any mistake with the highest priority and find solutions for strict mistake avoidance.
3. Everybody involved in production is obliged to improve products and processes continuously.

To implement these rules, a set of methods is proposed (frequently using Japanese language terms) that may be interpreted as patterns of lean production. The rules most frequently referred to are (not always strictly applied in the original *Toyota* sense):

- *Kanban* (use card), a material flow and inventory control procedure, supporting just-in-time production and minimizing lead time as well as inventory levels
- *Kai zen* (improve the good), prescribing the continuous improvement process (CIP)
- Six Sigma, calling upon statistical backgrounds to achieve waste ratios, measured in parts per million (ppm)
- The five S's: *seiri* (sort out), *seiton* (clean up), *seiso* (keep proper), *seiketsi* (make orders your rules), and *shitsuke* (improve all things continuously) as simple rules of behavior for everybody in the organization
- Management by view: all processes are visible and transparent in order to recognize irregularities immediately
- *Poka yoke*, technique to obtain error-free transformations and moves
- Quality circle (QC, TQM): readiness to improve the quality of all processes, executed by teams
- *Jidoka*, to stop process steps (including the impressive *pull the emergency line* option) if errors are detected/assumed

For the implementation of the methods and rules, ten steps to lean production are recommended:

1. Eliminate waste
2. Minimize inventory
3. Maximize flow
4. Pull production from customer demand
5. Meet customer requirements
6. Do it right, first time
7. Empower workers
8. Design for rapid changeover
9. Partner with suppliers
10. Create a culture of continuous improvement

Instead of devoting resources to extensive planning, the Toyota production system focuses on reducing system response time by enabling the production system to adapt instantly to market demands. All automobiles are made to order, resulting in a number of effects: delivery on demand, minimization of inventory, maximization of the use of multiskilled employees, flattening of the management structure, and concentration of resources where they are needed.

Lean production shows revolutionary attributes, providing an enormous productivity boost [15.104]. Therefore other companies (even competitors) have tried to adapt this methodology and integrate it into their corporate strategies, coining terminologies of *X*–production systems (where *X* is the company or brand name) that summarize rules and methods as mandatory parts of multisite company-wide standards (footprints).

15.7.2 Agile Manufacturing

Agile manufacturing [15.105] was developed from the synthesis of a lot of companies with individual abili-

ties or responsibilities, which come together in a joint venture. It was intended to be able to use the sum of abilities and resources of all the partners together. These joint-venture companies are called virtual companies, able to reshape and change quickly. The implementation of agility aims for intensive exploitation of organizational knowledge. For this purpose individuals are motivated to collaborate closely in dynamic teams focusing on clearly defined opportunities in the market [15.106]. Employees and information are the key factors that should ensure the superiority of agile companies compared with the surrounding competition.

Agile manufacturing aims to enable an increase in performance of companies by achieving major steps toward flexibility and time to market. Hence, the value chain has to be compatible with fast movement of products to market.

The main target of agile manufacturing is the development of durable success in an environment of continuous and nonpredictable change. For success the times of processes inside the value chain have to be in accordance with the fast movement of products to market. The agile approach aims to serve mass markets, but all individual wishes of the customers should be granted. Cooperation with other enterprises helps to produce new products faster and cut costs, reducing the risks for all partners. Usually joint ventures – vertical and horizontal mergers within the value chain – establishing virtual companies are the strategic aim.

A characteristic of this concept can be identified is the rapid change of the structure of the networked organizations. This includes all efforts to make use of the latest information and communication systems as well as the integration of fast reprogrammable technologies in production [15.107].

Within the agile concept the employee is the focus; the necessity for a high level of knowledge and education is clear. The employee must act as an entrepreneur in the company and must be actively engaged in all relevant processes, because of the low and simple level of organization. The dynamic development of a team actively helps to develop the creative and innovative talents of other members of the teams. Management assists by developing company culture, supporting creativity, and being open to experiments and risk-taking by the employees. Leadership is based on motivation, support, and especially on building trust.

15.7.3 Bionic Manufacturing

Recognizing that even so-called *flexible manufacturing systems* are not able to fulfil the demands of customer-specific parts production, the bionic manufacturing system (BMS) was developed [15.108]. BMS is an approach that aims to master the future demands of manufacturing systems through the application of technology that mimics the nature of living beings. The core idea of the BMS is the *creative system*, in which the materials provide the necessary information to the manufacturing equipment. Intelligent methods respond to this information using flexible and autonomous technical units.

BMS draws on the results of artificial life research, by embedding DNA-type information into the materials to be processed. Material with this stored information is passed onto the manufacturing system so the product can be built accordingly. Information is transferred to the operating manufacturing system, offering the following characteristics:

- Ability to learn and identify necessary tasks
- Ability for self-maintenance
- Ability for communication
- Ability for self-creation to acquire new product and technology knowledge [15.109]

The product collects further information and passes this on as recycled raw material back to the manufacturing system, which interprets the information again and develops further. The basic information (for example, the method to move = car) survives as DNA information, whereas the bionic manufacturing system information enables the system to develop independently and adapt to cultural and temporary demands.

This far-reaching vision has remained in fashion for a long time and currently miniaturized transponder technologies such as RFID are enabling powerful solutions of this kind [15.110].

15.7.4 Holonic Manufacturing Systems

Holonic manufacturing systems (HMS) may be envisioned as specific interpretations of BMS, as HMS are also able to adapt and incorporate new products, new organizational structures, and new technologies. Holonic organizations (or *holarchies*) support the setup of very

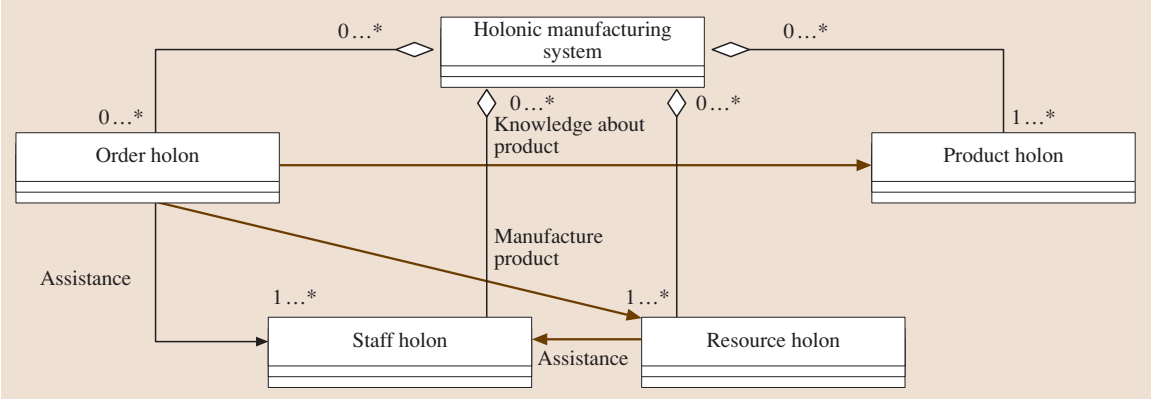


Fig. 15.41 Basic setup of a HMS

complex systems that are highly resilient to disturbances (both internal and external) and adaptable to changes in the given environment.

The basic ideas for this concept originate from *Koestler*, who identified behavioral properties of living organisms and social entities [15.111]. He stated that complex systems are adapted by evolution, where a holon (from the Greek meaning "whole", as coined by Koestler) is an identifiable part. The entity itself is made up of subordinate parts and in turn is always part of a larger whole. The self-reliance property ensures stability and abilities to cope with disturbances, as holons handle processes and problems on their particular level that do not involve higher-level holons. Holons may receive instructions and are therefore controlled (to a certain extent) by higher-level holons, ensuring the effective operation of the larger whole.

This idea can be transferred to manufacturing by introducing holonic manufacturing operations constituting a flexible manufacturing system Cooperating control units solve a common control problem by exploiting the self-reliance property of holonic systems, finally leading to what is known as *holonic manufacturing systems* (HMS) [15.112]. Progress in software engineering and numerical control allows full incorporation of holonic behavior into operations execution. The resulting HMS control product–resource–order–staff–architecture (PROSA) makes automated processes (the *shop floor* as well as the upper layers of the automation systems) more flexible by avoiding overheads on the control level [15.113].

Holonic development is still ongoing at the international scale. The HMS consortium has published the following list of definitions (among others) to specify the translation of holonic concepts into a manufacturing setting [15.114]:

- A *holon* is an autonomous and cooperative building block of a manufacturing system for transforming, transporting, storing, and/or validating information and physical objects.
- The capability of a holon to create and control the execution of its own plans and/or strategies is referred to as *autonomy*.
- The interaction of a set of holons to fulfil a common goal is referred to as *cooperation*. This will normally cover a process whereby a set of holons will develop mutually acceptable plans and will execute these plans.
- A system of holons that can cooperate to achieve an objective is called a *holarchy*. The holarchy defines the basic rules for autonomy of, and cooperation between, holons.

PROSA implements the HMS architecture by four main types of agents, implementing four main types of holons: the order holon, product holon, resource holon, and staff holon (Fig. 15.41) [15.115].

Based on the HMS approach and the PROSA findings, a number of holonic control architectures have been developed in recent years [15.116].

All architectures support steps toward more flexible manufacturing systems by additional implementation of interfaces on the field control level as well as to the ERP level.

15.7.5 The Fractal Company

A turbulent environment (market, surroundings, and resources) and acceleration (turbulence) of market shifts

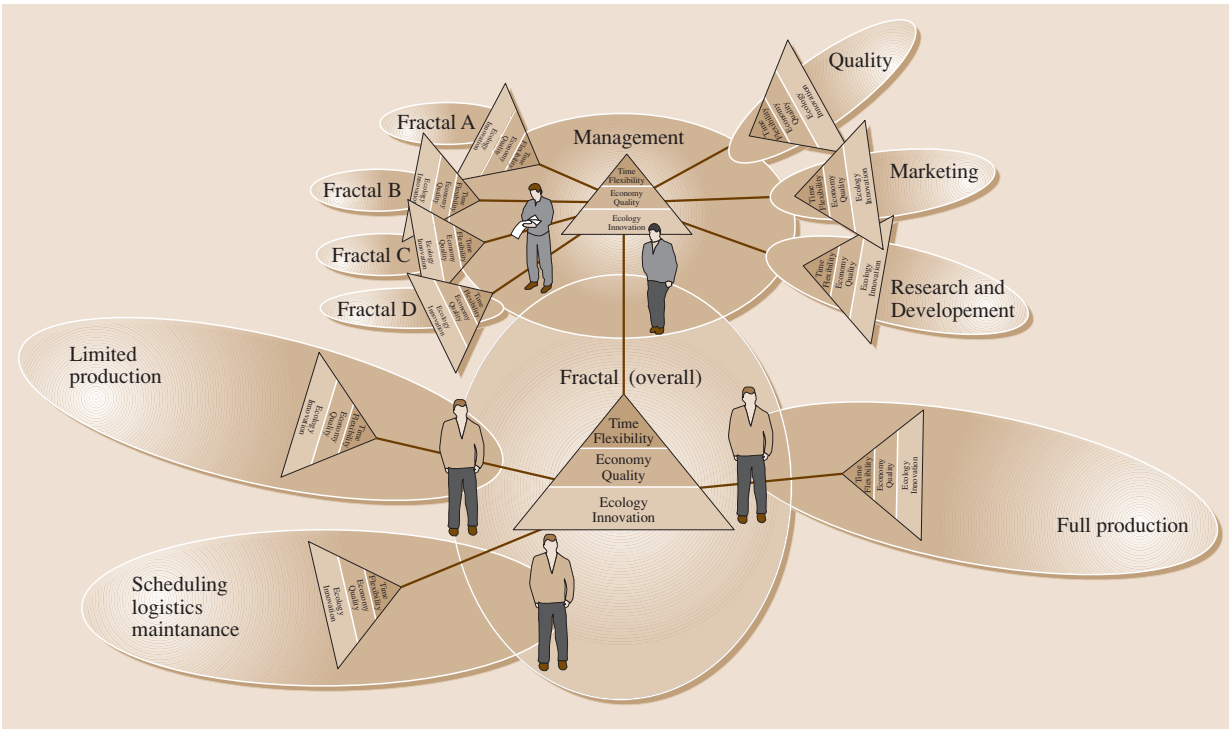


Fig. 15.42 Fractal company: A typical organization chart

lead to enormous complexity for companies. So, networks of autonomous units, forming versatile company structures, may be seen as an adequate organization structure [15.117].

Companies are decomposed (companies in the company) and transformed into networks of units [15.118]. In order to direct these networks of units or network organizations, additional instruments are engaged that master this complexity. For the management of complexity in geometry Mandelbrot proposed the fractal view, where complex geometrical structures and patterns are discovered as results of simple rules of self-similarity (e.g., Mandelbrot set, Koch's snow flake) [15.119].

This self-similarity property, found in geometrical structures, has successfully inspired organization theory [15.120], resulting in the fractal company concept [15.121].

The fractal company envisions enterprise organization as consisting of autonomous team units – so-called fractals. These fractals are attracted by (internal as well as external) market opportunities. Thus the companies' market opportunities are to be taken by the units directly; therefore the strategy parameters have to be *translated* into the units' language. The companies' objectives are self-similarly broken down into units' objectives to *guide* the teams (i. e., the fractals) [15.122].

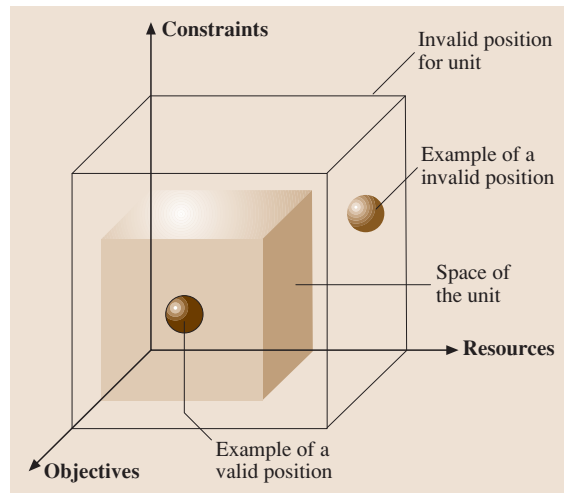


Fig. 15.43 Space of activity (SoA) by mappings for planning and control of production networks with position cases

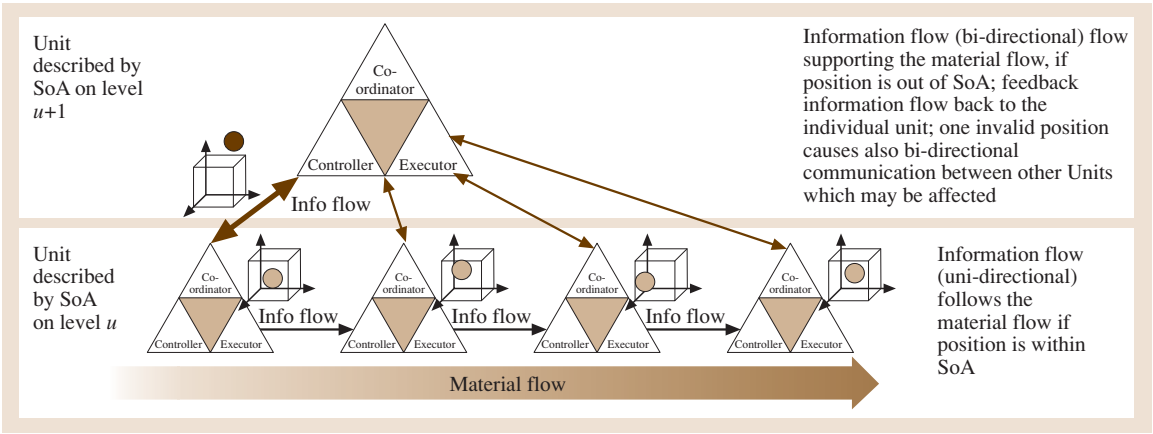


Fig. 15.44 Information flows for harmonising goal settings/SoA volumes caused by invalid position of higher level SoA

The fractal company is therefore defined as an organization consisting of goal-orientated self-similar team units – the fractals – that are described by the assigned objectives and outputs (Fig. 15.42) [15.123].

Resulting properties are structural versatility, dynamics, and vitality. Operating a fractal company is a continuous development, an objective alignment and improvement process. As the most important consequence, detailed job descriptions and schedules have to be abandoned and replaced by self-organization, visualization of the objectives' updates, and increased decision power of the employees.

Decision-making in network organizations such as the fractal company largely involves negotiation and communication between units, since there is no central control [15.124]. The decision support model used for the fractal units is the space of activity (SoAs) (Fig. 15.43), which describes the units' activities and success.

All units' objectives are to be harmonized with the overall company objectives. Therefore, consistency procedures for the networked company organization must be applied (Fig. 15.44). SoA parameters are the input for the decisions about maintaining the self-organization mode or calling for stronger company

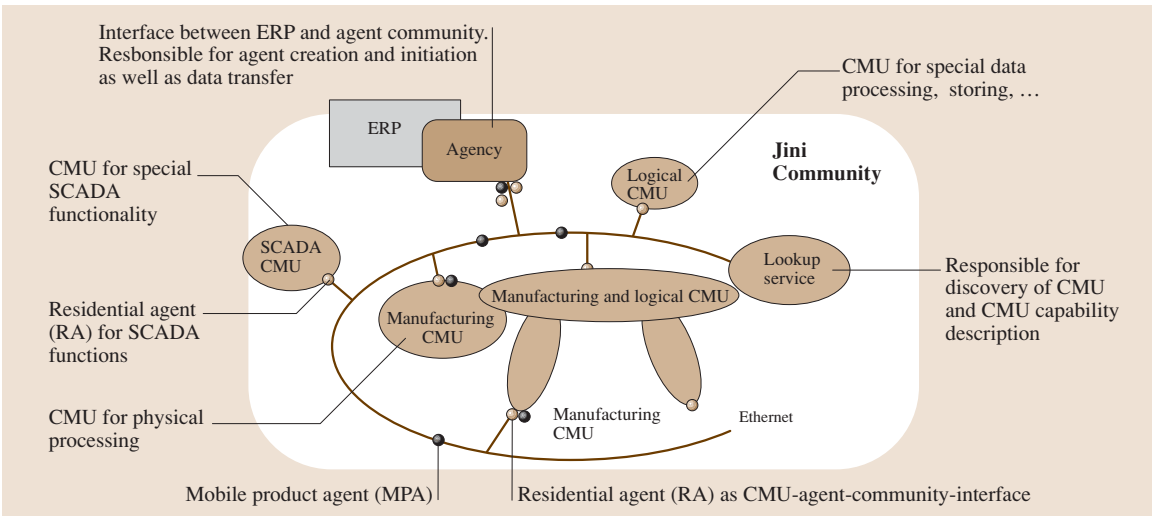


Fig. 15.45 PABADIS control architecture (Jini – Java intelligent network infrastructure; SCADA – supervisory control and data aquisition; CMU – cooperative manufacturing unit)

Table 15.11 Six key attributes of advanced organization concepts

	Fractal	Holon	Bionic	Agile	Lean
Origin	Geometry	Physics	Biology	Nature	Theory of Constraints
Analogy	Fractal	Molecule	Organism	Agility	Mechanics
Key principles	• Self-similar • Self-organization • Self-optimization • Process-orientation • Vitality and dynamic	• Self-similarity • Autonomy • Distributed intelligence	• Autonomy • Cooperation • Self-organization • Self-optimization	• Enrich the customer • Master change • Resources • Cooperate to compete	• Perfect first-time quality • Waste minimization • Continuous improvement • Pull processing • Flexibility • Long-term relationship with suppliers
Methods of structuring	• Product structure • Equipment • Employees • Flow of material	• Tasks and their similarity	• Knowledge base of autonomous elements	• Potential of supply and demand	• KAIZEN • POKA YOKE • KAN BAN
Objective	• Cohesive organization structure along the value chain • Minimizing interfaces	• Mutability • Ecologically sensitive • Robustness	• Dynamic adaptation and organizational ability to learn	• Adequate flexibility of the production system to react to market changes	• (MUDA) (Japanese for 'eliminate waste'
Appearance	• Divisions with variable organization structure	• Event-driven communication between machines	• Sender–recipient relationship between resources and material	• Flexible organization structure • Interdisciplinary teams • Flat management structures	• JIT • Pull • Value stream
Configurable resources	• Employees • Divisions	• Equipment	• Elements with inheritable knowledge (equipment) • Elements with learnable knowledge (material)	• Employees • Production and information technologies	• Objects • Equipment
Development of resources	• Skills and social competencies	• Distributed systems	• Evolution of knowledge	• Syntheses of resources	• Best-practice benchmarks

interference [15.118]. If objectives of an unit are not achieved, the company top level must be involved: However, depending on the unit's ability to profit from the opportunities in dynamic environments, the company may lower its influence on the unit, allowing or even supporting autonomous activities of (self)-optimization, (self)-organization, and (self)-structuring.

Figure 15.44 illustrates only one case of the companies' meshed control loops established by SoA interactions or interferences. Higher structure levels of the company are represented by SoAs, self-similarly containing all corresponding SoA structures. Increase of market complexity (uncertainty, turbulence, and unpredictability) may force the company to expand the spaces of activities. More-foreseeable steadier conditions reduce complexity. Such conditions make smaller SoAs more effective, in the limit degenerating to a point a the origin for uniform mass production. The SoA model can therefore be envisioned as a generalization of the job definition in the hierarchical organization.

Based on the SoA setup as outlined, negotiation and decision procedures between the units can be formalized and assigned to software agents. Making use of internet technology as well, a next-generation control architecture aiming at adaptable manufacturing equipment, flexible integration of different types of control systems, and distributed control devices for the execution of manufacturing orders can be derived: e.g., the plant automation based on distributed systems (PABADIS) architecture [15.125]. A PABADIS-compatible control system consists of four components (Fig. 15.45): the agency, the agent community, the CMU community, and the lookup service [15.126].

The main benefits of PABADIS are improved order and resource flexibility and the generic interface structures of the control units – the control building blocks. The integration of a new order or a new product only requires the specification of the related set of manufacturing process data. Product and order agents are generic and therefore valid for each possible application case. This avoids the implementation of new agent types when changes occur.

These properties and the internet compatibility of the PABADIS architecture has attracted the attention of leading vendors and users of planning and control systems for manufacturing. Within the PABADIS'PROMISE approach the PABADIS architecture will be improved and extended to all levels of control [15.110].

15.7.6 Summary

Fierce competition and saturated markets have forced enterprises to cope with the diversity in customer requirements and the speed of the demands of the markets. Organizations had to become more flexible, versatile, adaptable, and reconfigurable in structure.

It is evident that this paradigm shift has generated radical innovations in organization. Of the numerous progresses, the most important generic concepts have been outlined in this Section. These approaches are characterized and compared in terms of their key attributes in Table 15.11.

15.8 Interorganizational Structures

Individual organizations are characterized by a mission, which states the scope and principles of their business. A mission tends to have a restricted focus with respect to market and product/service, even though, particularly for large corporate companies operating in different market sectors and product/service areas, the mission statement may consist of a general statement of intent.

Throughout their evolution, companies continuously face new challenges derived from the evolution of products, technology, and markets, and face the need to acquire improved governance capability over the market and the product/service lifecycle, while being compelled to cope with limited resources in terms of finance, disciplinary competence, capacity, and risk sustainability.

Without addressing issues associated with monopolies and vertical trusts, this increase of control can be pursued either through internal growth of a company (possibly using merger and acquisition levers), or by seeking cooperation with external entities (Fig. 15.46).

Further progresses in ICT, especially the maturity of the World Wide Web (WWW), offering as-yet-unknown possibilities to transfer/store huge volumes of information throughout the globe, accelerates further decentralization. Dispersed production units may cover product segments, functional units or entire production stages, dissolving boundaries within and between

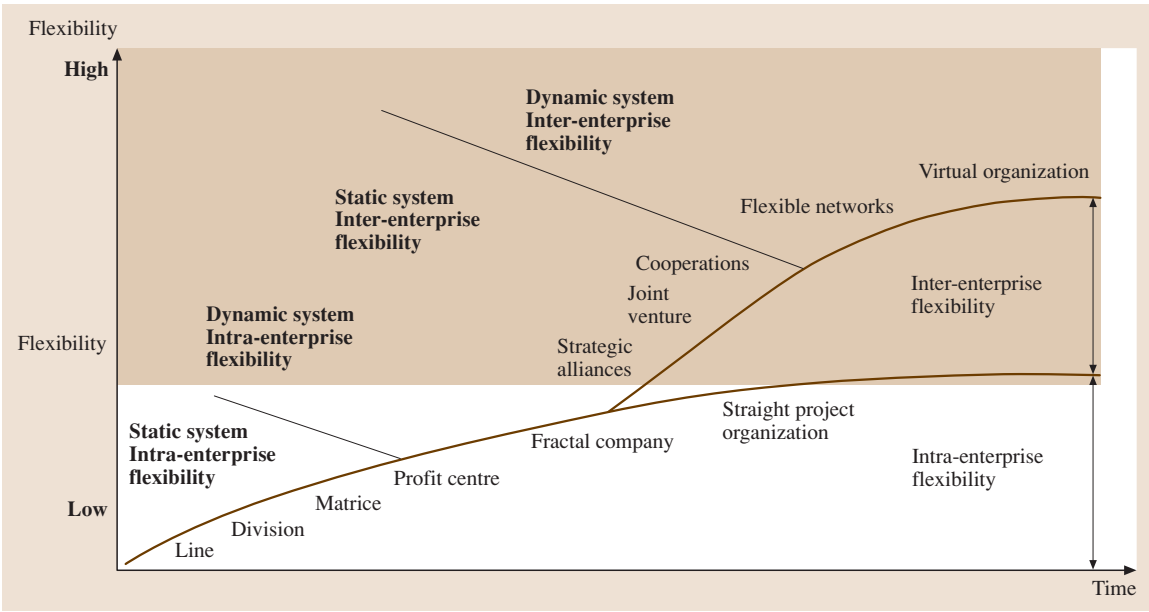


Fig. 15.46 Evolution of inter-enterprise flexibility (after [15.14])

organizations/companies completely. Such distributed structures, known as virtual organizations (VO) or extended enterprises (E2), are able to develop and deliver products and services faster and cheaper. Benefits result from combining complementary capabilities and experiences which would otherwise need to be acquired at high cost. Broken down into equipment units, objective fulfilment is executed by globally distributed networked operations resembling a production grid of communicating units.

15.8.1 Cooperation

Cooperation [15.127] is the practice of working in common, with commonly agreed-upon goals and possibly methods, instead of working separately in competition.

Cooperation mechanisms have been the subject of several studies in the area of theory of games [15.128, 129], providing theoretical support for the identification of behavioral principles both in the launch of cooperation and in the management of its operation.

Depending on the relationships among the collaborating entities, cooperation may take different legal and organizational structures:

- When there is a clear lead by one of the participants, cooperations are shaped and managed according to hierarchical schemes (Fig. 15.47), typical of the supply-chain environment, with organizations that

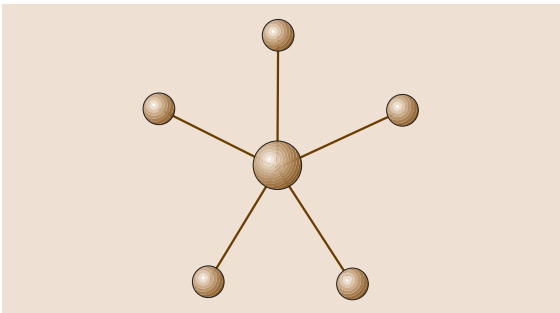


Fig. 15.47 Dominated co-operation

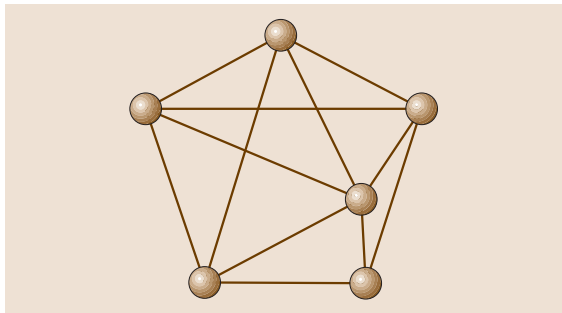


Fig. 15.48 Co-operation among equals

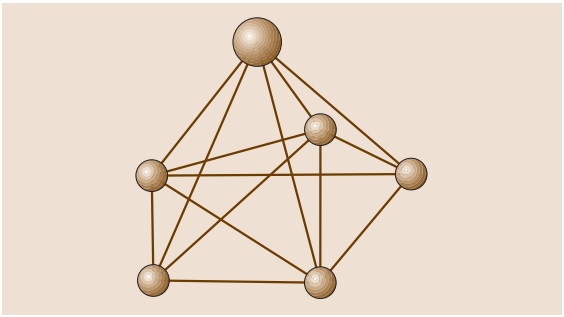


Fig. 15.49 Extended enterprise relationships

tend to be based on contractual agreements and normally resemble the work breakdown structure (WBS) defined and set for the achievement of the collaboration goal.

- In the case of cooperation among equals (Fig. 15.48), organizations are shaped using a wide degree of freedom, depending on applicable law and on the scope of the collaborative initiatives.

Cooperative structures have a quite wide variety, and identified classes are not mutually exclusive, as some blending may occur, as in the case of extended enterprises (see below), where dominated cooperation is mixed with cooperation among equals to allow for peer-to-peer cooperation among nonleading partners (Fig. 15.49).

In the following, reference will be made to six classes of interorganizational structures (Fig. 15.50):

- Alliance
- Network
- Supply chain
- Virtual organization
- Extended enterprise
- Virtual enterprise

which overlap and may actually be a substructure of wider types, as in the case of virtual enterprises with respect to alliances.

15.8.2 Alliances

The term *alliance* is a generic term used in business to refer to several kinds of cooperation frameworks, mostly characterized by a *peer-to-peer* partnership condition.

Alliances mostly respond to the need of individual companies to minimize the risk in initiatives that may

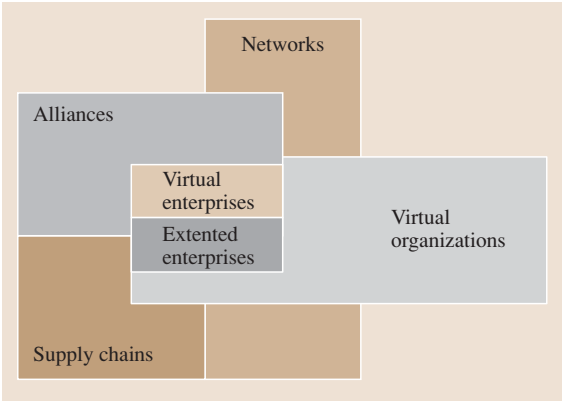


Fig. 15.50 Co-operation structure types

substantially impact the health of the organization in case of failure.

Three main drivers have been identified [15.130] for the launch of an alliance:

- Globalization: the reduction of trade barriers and subsidies pushes toward larger scale and international presence.
- Time to market: speed of competitors is growing, and traditional organic growth cannot keep pace with it.
- Diversification: companies tend to open a second, third or fourth business to spread the risks of investors in company's stock. The idea is that, by entering a business that does well in recession of other company *legs*, the stock price and cash flows will hold up during business-specific contingencies.

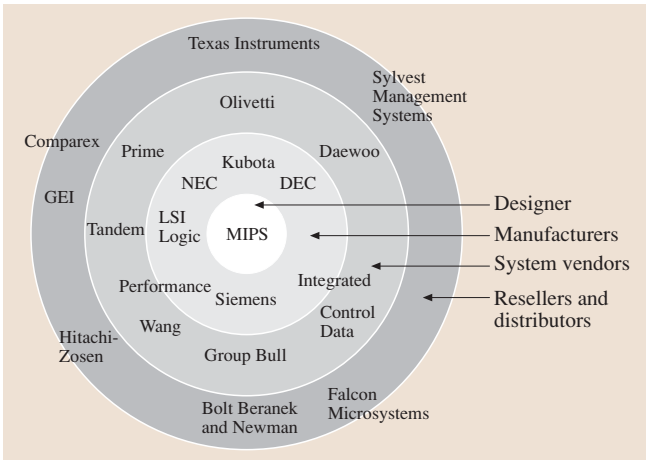


Fig. 15.51 The MIPS RISC constellation

An additional consideration is cost to market to develop new products or services at lower risk and shared risk to retain market share, where market share is traded against development cost, or to penetrate new markets, possibly by working with a recognized partner in a market segment to open opportunities at lower risk and cost.

When alliances involve several partners, it is usual to refer to them as *constellations*, as in the example of the alliance setup by MIPS in the early 1990s for the development of the innovative reduced-instruction-set computer (RISC) technology (Fig. 15.51).

Alliances imply the joining of companies to pursue common, shared objectives, tend to have a limited number of partners, and are usually associated to particular kinds of cooperation:

- Alliances are the mechanism used to achieve results that would not be within the reach of individual entities, or in general to minimize risks and organizational costs of particularly challenging or expensive initiatives.
- An alliance may also be a safeguard mechanism through which otherwise competing entities perform a high-risk activity together so that failure of the initiative cannot lead to a strong competitive edge for nonparticipating parties (as in the development of radically new technologies).

Sometimes alliances represent a real alternative to mergers and acquisitions, to enable greater market competitiveness, as in the case or commercial airlines, which enter code-sharing agreements to improve their market reach and their capability to satisfy the requirements and expectations of their own customers. In some cases, alliances provide a trial environment for mergers, allowing for testing and tuning of evolved, integrated business processes for the new integrated company. Conversely collaboration on specific products and services allows organizational cooperation without the risks and inherent costs of merger or acquisitions.

From the point of view of management issues, the establishment of specific collaborations schemes with external entities meets several objectives:

- Keeping the focus of interaction on specific business objectives
- Allowing for the control of advancement and efficiency of activities
- Improving the interaction capability between different entities, focusing on joint strengths in the alliance
- Achieving cost-effective routes to market

Alliances can be managed through joint legal entities, i. e., separate companies with stock-holding shared among the partners, or on a pure contractual basis (typically assuming the contractual form of *joint ventures*).

While in any case partners maintain their individual business focus, they allocate resources to the common initiative, and tend to establish an independent management office so to ensure that agreed mutual obligations and business focus is maintained throughout the life of the alliance through a central point of control, communication, and responsibility.

The business concept for the alliance, particularly when in the form of joint creative developments (as in the case of alliances for the exploration and development of radically new technologies), may change over time, and the existence of an independent management entity helps to identify the most appropriate organization evolution strategies that are capable of accommodating the needs and expectations of all partners.

Whether through actual or virtual collocation, human resources are usually allocated to alliances on a full-time basis, as in the case of large projects in integrated companies. This is likely to raise several issues in partners' personnel management strategies associated with the possible loss of feeling of belonging by deployed personnel and to frequent problems in the management of professional careers and the difficulty of ensuring adequate recollocation of personnel within partner companies at the closure of joint operations.

Even in the case of alliances that may lead to substantial benefits to partners and to significant evolution of core business (so-called *strategic alliances*), partner companies maintain their individuality. They are driven [15.131] to frame their cooperative initiatives within the global company strategy, which may well include an *alliance strategy* as a policy to achieve company vision and goals, but which cannot be substituted by a single strategic alliance, as that would de facto result in the same level of risk for the company as before, thus missing out on the main benefits yielded by the alliance.

As a final note, it is worth quoting the top ten factors identified by *Benjamin Gomes-Casseres* [15.132] for the success of an alliance:

1. Have a clear strategic purpose: treat alliances as tools of a business strategy.
2. Find fitting partners: select partners with compatible goals and complementary capabilities.
3. Specialize: allocate tasks and responsibilities in accordance with each party's best capabilities.

4. Create incentives for cooperation: the cooperative attitude must be nurtured, particularly when partners were formerly rivals.
5. Minimize conflicts between partners: allocate roles in such a way to avoid pitting one against the other in the market.
6. Share information: communication develops trust and keeps projects on target.
7. Exchange personnel: personal contacts and site visits are essential for maintaining communication and trust.
8. Operate with long time horizons: sharing a vision strongly supports the solution of short-run conflicts.
9. Develop multiple joint projects: successful projects can help when other projects enter critical phases.
10. Be flexible: alliances are open-ended dynamic relationships that need to evolve in pace with their environment and in pursuit of new opportunities.

Those factors imply an underlying company strategy based on time to market, cash flow, investment, the availability of resources, and the capability to manage alliances in line with organizational investment cycles for new products and the natural lifecycle of products to secure maximum revenue.

15.8.3 Networks

Compared with alliances, networks provide a looser kind of relationship [15.133], as they imply multiple close but nonexclusive relationships, whereas alliances imply the creation of a joint enterprise, at least over a limited domain.

Networks generally exist on the basis of complementarities and potential synergy among members, which lead joining companies to associate with the network a greater opportunity for acquiring and retaining a competitive edge in the market:

1. To reduce uncertainty: the relationships developed through the network allow for avoiding the uncertainty associated with impersonal, nonrepeatable, and purely exchange-based market transactions.
2. To provide flexibility: the network provides a greater expectation of immediate resource reallocation, and allows for looser constraints in the establishment of joint product/service-specific initiatives.
3. To provide capacity: the likely availability of individual network members' spare capacity allows companies to reliably pursue opportunities that lie beyond their own contingent capacity while improving the management of contingencies in day-by-day operations. Furthermore, the network allows for the rapid provision of resources and skills which are lacking in a company through access to qualified resources from other known members.
4. To provide information: members of networks have easier access to industrial intelligence, as open information constitutes one of the mechanisms for establishing relationships and one of the principal reasons for members to join.

A network is expected to offer to its members relationships that can lead to profits in the future. Based on this motivation to join, the power of a member has [15.134] five main sources: economic base, technologies, range of expertise, trust, and legitimacy, i. e., the characteristics that are most likely to support the success of initiatives by other members.

Such expectations lead to governance principles that are close to historical socioeconomic organizations founded on reputation (both individual and acquired through belonging to groups of qualified members, such as a family or a clan) and on the provision of services and favors, which do not offer immediate reciprocity but which are aimed at strengthening relationships in view of future opportunities.

The networking principle is applied as a mechanism for both dominated cooperation and for cooperation among equals. Cooperation domination is typical of advanced supply chains and extended enterprises (see below), whereas *peer-to-peer*-based networks constitute a characteristic dynamic entity which needs specific management provisions to be maintained.

The peer-to-peer network is not a substitute organizational form for the integrated firm, even though collaborative business operations are usually conducted in accordance with classic customer–supplier relationships; roles are nonetheless dynamic and vary with business opportunities, and depend on contingent conditions that are mainly related to commercial positioning, competence ownership, dimension, and geographical location.

As there is no stable leadership in these structures, the relationship glue tends to vanish with time, even for networks which are strongly focused on specific market sectors or composed of members with close geographical locations. There is therefore a need for a *network entity*, which is in charge of representing the network identity and of nurturing network relationships, typically through the provision of shared services at the technical, technological, commercial or legal level.

Challenges that must be faced in networks are those typical of closed systems: network relationships may evolve towards the exclusion of newcomers and the limitation of the search for partners to the group of network members, with overemphasized trust that might cause ineffective global performance when compared with a supplier search in the wider market. On the other end, too loose relationships may actually render the network itself ineffective and closer to a *club* than an efficient business organization, even though some national regulations, as in the case of Italian Districts, may nonetheless prize formal membership to such a network.

15.8.4 Supply Chain

The complexity of modern products and services is rapidly growing, and no company can afford to manufacture its product efficiently and fully alone. During the development phase of a product, each part is analyzed for a *make or buy* decision, and an increasing number of components (both material and immaterial, such as analytical services) are procured/sourced from other companies. A similar approach can be taken for the development of a service by either using internal capability or outsourcing where more cost effective, or where a specific capability or capacity is needed to achieve the offering. Furthermore, there is currently a trend towards company downsizing, where internal noncore functions are replaced by subcontracted providers, often constituted as an independent company through the outplacement of salaried workforce.

The sequence of suppliers that contribute to the final product or service (both in direct and overhead activities) is usually called the *supply chain*.

The configuration of the supply chain for a product or service is the result of a selection among alternative suppliers, performed in accordance with a strategy that is certainly based on price considerations, but must also account for several characteristics derived from the overall company strategy:

- Qualification: suppliers are critical for the final result, and therefore need to enjoy a sufficient level of trust in their capabilities; furthermore, specific compliance to regulations and standards is often required for various market and product sectors, and such requirements must be met by all suppliers. The qualification process is normally long and expensive, and must be repeated over time, both to re-register the supplier against currently procured capabilities, but also to consider its possible new capabilities and its potential for evolving technologies and business processes.
- Capacity and stability: the size of suppliers and guarantees of continued operation over the lifecycle of a product allows companies to avoid planning investments for the allocation of internal resources and facilities.
- Reliability and performance: the quality of supply and its timeliness, whenever possible combined with superior knowhow deriving from specialization, together with a consistent level of performance and with the provision of robust results allow for smooth management of operations by the contractor, and minimizes the risks associated with fractionated development and manufacturing.

In the supply chain, responsibility and risk for the final results remains with the lead company, so that interaction among supply-chain participants is typically hierarchical, with relationships between entities that are associated to the product and work breakdown structure that are consequently classified into sequentially ordered tiers.

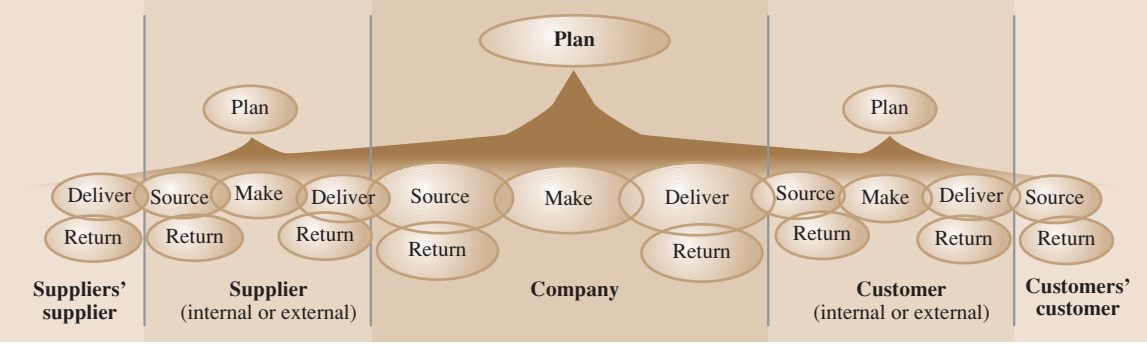


Fig. 15.52 The SCOR processes [15.135]

It is becoming frequent in modern supply chains built for particularly complex products, such as aircrafts, that the manufacturer and first-tier suppliers enter into risk-sharing agreements that provide the manufacturer with additional financing sources in the development phase and reduce the risk that it must sustain; to balance this additional risk, suppliers require higher visibility on strategic decisions and on the reliability of market projections, which tends to strengthen the relationships and the mutual trust, at least for upper-tier levels.

The manufacturer keeps in any case the role of architect, which implies responsibility for the specification of component requirements (as a minimum at a functional level) and of boundary and interface requirements.

The supply chain is then governed through contractual agreements/orders, which specify the scope of provisioning and the conditions for supply (planning and scheduling, delivery methods, invoicing and payments, etc.).

Management of the chain is consequently driven by contractual provisions, with communication flows that are usually associated with expediting and advancement reporting, as well as on compliance of deliveries with requirements.

Operations in a supply chain and its effectiveness [15.136] can be evaluated through its modeling in accordance with the supply-chain operations reference (SCOR) approach. This methodology was developed in the late 1990s by the Supply Chain Council to model and evaluate supply chains. The value chain is described as a sequence of standard processes, namely the *make*, *source*, *deliver*, *plan*, and *return processes*. The contribution of each participant in the value chain can be described by at least one of these processes, which leads at the top level to the general architecture of the supply chain. Figure 15.52 provides an example of how SCOR describes a supply chain. This top level can be specified at the level of process categories (level 2) and detailed process elements (level 3). It is obvious that SCOR also focuses on supply chains to describe and monitor existing value chains with the objective of achieving optimization.

The benefit of the SCOR model is that it provides standardized processes which allow one to model the whole intercompany value chain with a single method – if, of course, all network partners agree to this standard.

15.8.5 Virtual Organizations

The term *virtual organization* appeared in the 1990s to qualify groups of companies that establish cooperative operations through the extensive use and exploitation of new information and communication technologies.

This classification does not imply a specific kind of relationship, so that both dominated cooperation and peer-to-peer cooperation are included.

In particular, when the paradigm of virtual organization is applied to supply chains, the expression *extended enterprise* is used, whereas for virtual organizations among equals it is common to use the terms *virtual corporation* or *virtual enterprise*.

The paradigm of virtual organizations responds to the need to improve the effectiveness and efficiency of cooperation among multiple organizations. In fact, unless the scope of the collaboration is particularly simple, the interaction among independent organizations (and possibly individual professionals) results in a huge number of problems, resulting from the need for harmonization of cultures, processes, and policies.

Virtual organizations leverage the capabilities of ICT to enforce advanced work methods in the partners' organizations, with the aim of establishing operations that use all mechanisms available to the vertically integrated enterprise, while maintaining the greatest level of flexibility in terms of capacity and capability.

The role of ICT is particularly relevant for ensuring team working through virtual collocation, integrated management of development and manufacturing planning, rapid administrative management and efficient cost controlling, and information sharing as well as knowledge management.

This capability allows virtual organizations to have geographically dispersed partners which are selected to have the best capability as required by the scope of the organization, compatible with existing constraints, and pursuing maximum mutual profitability.

Due to their openness and dynamics, virtual organizations are particularly subject to the issue of trust creation and maintenance, unless they can leverage pre-existing relationships as in the case of partners belonging to the same network.

Special care is needed in particular for the management of knowledge in virtual organizations, as the increase of communication flows requires that knowledge is more freely exchanged among participants. Nonetheless, one of the most beneficial characteristics of this type of structure, i. e.,, its capability to evolve

dynamically by changes in partnership according to performance requirements, may be a restraint to the willingness of individuals and organizations to open up some of their knowhow. From this viewpoint, it is important to note that this issue is less critical for product sectors that undergo rapid technological changes, thus rapidly making knowhow obsolete, whereas market sectors with mature technology are confronted with greater reluctance to exchange knowledge due to its relevance for the power, positioning, survival, and progression of the individual and company involved.

However within supply chains and partnerships where the exchange and interaction of knowledge is fundamental to business processes, it is also important that the knowledge available and applied within each organization is comparable, complementary, and well maintained.

Such criticality is reflected in the degree of formalization and standardization of interactions in a virtual organization, which provide support for coupling and decoupling partners with ease.

From the work developed in the European project ARICON (standardised assessment of readiness and interoperability for cooperation in new product development in virtual organization) [15.137], a set of six critical areas (Fig. 15.53) are identified as prerequisites for the success of virtual organizations devoted to new product development:

- Business models and strategies: the sharing of common goals and vision for the cooperation, with compatible business models among the partners is the basis for motivation and effectiveness in decision making.

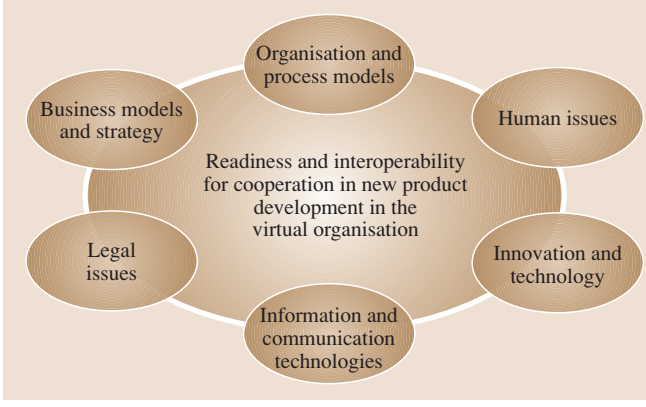


Fig. 15.53 Critical areas in virtual organizations

- Organization and processes: roles, responsibilities, and communication lines to be transferred into operational ICT must be in line with harmonized processes and well-specified interfaces.
- Legal issues: multinational organizations need to have clearly established behavioral rules and contingency management policies and contractual relationships.
- Human issues: complementary, regulated, harmonized partner companies' cultures, individual motivation, professional growth, and operational and motivational training provide a common framework for efficient collaborative work and the option for personnel exchange opportunities for business reasons, cultural exchange or personnel development purposes.
- Technology and innovation: complementarities and alignment among partners' technologies ensure the capability to develop, manufacture, and maintain common products and services.
- Information and communication technology: the basic capabilities for correct organization's working fall in the ICT environment, which should ensure interoperability among the partners' systems and effective collaborative functions.

15.8.6 Extended Enterprise

The extended enterprise is an evolution of the supply-chain environment, and is characterized by the propagation of technical methods, tools, and processes along the supply chain.

The principle of the extended enterprise is harmonization of processes and practices by the propagation of a client's standard way of working towards its suppliers. This kind of harmonization tends to be substantially embedded in the information technology framework that is adopted for the operations, which in the case of the lack of de facto standards is normally specified by the extended enterprise leader, which applies a predefined organizational strategy.

The increased participation of suppliers is evident in the practices of co-design and risk sharing. Co-design, at the relevant level of responsibility of each actor, is the practice of participation in the development of a product within the supply chain. Such an approach is already in use in most complex chains, but with the limitation of physical collocation of personnel, so that only a small part of the development activities could actually be subject to co-design, and therefore with a specific focus on the early phases,

where deviations and misunderstandings may have a larger impact on final cost and time to market. The ICT infrastructure of the extended enterprise allows the collaborative phases to be extended through the life-cycle, resulting in a substantial improvement in effort alignment.

Due to the increased responsibility allocated to suppliers through the process of co-design, extended enterprise leaders tend to adopt a risk-sharing strategy to ensure organizational motivation for best performance. While being enforced through legal agreements, the additional risk taken on by suppliers is enacted through greater openness of strategic information, which flows from top to bottom of the hierarchical chain of suppliers.

The improvements achieved through the extended enterprise concept are therefore related to the establishment of common methods, reflecting harmonized interfacing at the boundaries of individual supplies and leading to:

- Increased ease of communication in the hierarchical path
- The capability to allow for cross communication among suppliers to improve cooperation capability

From the point of view of performance and cost, extended enterprises tend to leverage digital procurement systems not only for the exchange of administrative information during operations, but also for bidding and auction phases, even though such practice is applied with particular care for supplies that cannot be classified as commodities, due to the relevance of co-design contributions.

Thanks to the degree of digitalization and to the increased information flows, the extended enterprise is expected to perform substantially as an optimized single integrated vertical enterprise, with significant improvements in the capability of:

- Improving reaction time to contingencies
- Simulating and planning operating conditions
- Monitoring and controlling flows of materials and associated product data.

15.8.7 Virtual Enterprise

The virtual enterprise (often referred to as a *virtual corporation* [15.133]) is a particular form of virtual organization that is characterized by the precision of its scope/mission and by the actual creation of an organizational structure that can be substantially independent of original organizational structures in the participating companies.

The enterprise is typically composed of companies that provide different functions, in accordance with their capabilities and individual strategic goals.

A virtual enterprise constitutes a *single electronic business entity* [15.133], without a *physical* structure.

Although the enterprise is presented as a real company, it is not typically associated with a legal entity, as outlined in the European research project ALIVE (advanced legal issues in virtual enterprise) [15.138], where the contractual framework of *joint ventures* was found to be most appropriate for this kind of cooperation. The partnership composition of a virtual enterprise is essentially dynamic, to better fit evolving requirements and market contingencies; the greatest benefit of this feature is to ensure that the competitiveness conditions sought at the establishment of the enterprise are maintained throughout its life, but calls for specific formally agreed policies to ease the replacement of partners due to the *democratic* character of the enterprise.

A virtual enterprise is not dominated by one partner, but is a cooperation among equals, with governance systems that must ensure equal participation in strategic decisions.

The virtual enterprise organizational structure is orthogonal to the partners' organization; by its very nature, the virtual enterprise's organization and process can be ideally fitted to the product/service to be developed and/or provided to customers; nonetheless, due to the enterprise's temporary nature, it is not possible to establish and empower hierarchical relationships, and personnel hierarchical dependencies and power at partner companies must be reflected in the roles and processes of the cooperative corporation in order to prevent conflicts due to the allocation of responsibility for tasks.

The constitution of a virtual enterprise is usually associated with the identification of a business opportunity by a company which does not have the capability individually to pursue such an opportunity, the equivalent of the role of the *architect* for networks [15.139], i. e., the entrepreneur and possibly the provider of the inspirational vision.

Based on the opportunity identified a new role has been introduced for virtual enterprises: the business integrator [15.140], i. e., the role devoted to:

- Selection of partners
- Setup of governance structure, methods, and tools

Table 15.12 Comparison of integrated and virtual corporations [15.133]

Organizational dimensions	Integrated corporation	Virtual enterprise
Organization structure	Formal and flexible	Flexible network, flat
Decisions	Ultimately by fiat	By discussion and consensus
Culture	Recognizable, encouraging employees to identify	Pluralist, linked by overlapping agendas
Boundaries	Clear *us and them*	Variable
Management	High overheads	Minimal overheads
Power	From the board, ex officio	Through possession of competences in demand; being the brand company

- Guidance to align internal partners' processes to the operation of the virtual enterprise
- Management of operations for the enterprise

More than in other types of cooperative work, the virtual enterprise derives its effectiveness from the motivation of the partners, which is usually founded on a shared vision and complementary individual goals, and from a high level of trust and confidence among the partners. This is why networks are the natural breeding environment for virtual enterprises, which can work on existing relationships, whereas specific methods have been developed [15.137] to assess objectively the readiness of potential partners to join specific common business opportunities.

More so than for extended enterprises, in which (at least) the *systemic* knowhow for a product is held by the single leading organization, virtual enterprises show criticalities in the management of pre-existing and generated knowhow that must be mastered to allow for dynamic reconfiguration of the partnership as required for the best chance of success.

To summarize, it is worth reporting the table proposed by *Child* [15.133] for comparison of integrated and virtual corporations (Table 15.12).

15.9 Organization and Communication

Organizations constantly exchange information. By means of their manifold interactions, information and communication become determining factors in their competitive performance. Organizations that learn create more knowledgeable workforces for operations and process innovation, generating highly flexible organizations in which people will adapt to new ideas and changes through shared visions.

15.9.1 Terms, Definitions, and Models

Every kind of successful development of a company is based on the intra-organizational acquisition of knowledge, the development of knowledge, and its economical realiation. Basically, knowledge can be characterized by two features. On the one hand, knowledge represents information provided directly by the brain, or that can be retrieved quickly from information memories available to human beings. On the other hand, knowledge is related to the context of the acquirer. People must be able to display their information directly by means of a model-like illustration of realistic relations, conditions, and procedures – referring to means of intermediation (i. e., language) – in such a way that the information is purposeful and can be applied in a practice-oriented context.

The obtainment of knowledge by means of purpose-related handling and application by employees occurs in two fundamental acquisition modes:

- Original acquisition through one's own experience
- Derivative acquisition of knowledge through communication

In this respect, communication comprises all relations and tools of communication both within and outside a company, serving as the interaction with the system's surrounding or further units within the enterprise. It is characterized by the influence that one organism exerts on another with the intention of purposeful and targeted

exchange of information through a common interaction space (operational task).

The amount (and also the intensity) and quality (and also content) of possibilities for communication are mainly determined by the organization model in the enterprise. In addition to various structuring criteria resulting from different types of activities (depends on the production program) and its *interrelations*, this often depends on technical, technological, and economical necessities (utilization). Only in recent times have communication-oriented criteria for production gained significance, due to the increase of more-complex products and production programs in conjunction with the parallel extension and increased development of technology.

The foundation of communication is information, which is defined as a purposeful extract of applicable issues, aiming for the achievement of a target or the solution of a problem. The amount of information – clustered and purposeful, specifically directed towards an issue, processed in such a way that it will be generally accepted, checked and transferable – becomes knowledge. Furthermore, nonpurposeful, latent issues are also labeled as information since this kind of information can at any time be combined or condensed into knowledge by means of purpose-related references. Information therefore constitutes an image of the environment surrounding us and the various components and procedures used by humans to orient themselves in the world and to recognize and influence it [15.141].

Information as the basis for communication and the development of knowledge linked with it can be standardized according to different features and consequently evaluated by form-oriented methods. Some of the most important features are:

- Derivation (internal or external information)
- Degree of formalization and structuring (formal and conditioned, or informal and unconditioned information)
- Condition (primary or secondary information)
- Area of embodiment and application (special or general information)
- Quantifiability (quantifiable or nonquantifiable information)
- Programmability (programmable or nonprogrammable information)

The degree of formalization of information is essential for the internal embodiment of the information process and, accordingly, for the reliability of information processes in the course of operations. Structured information and its formal, i. e., organizationally ruled, course of communication includes all communication procedures, which by means of structuring, purposefulness, and repetition are related to the entrepreneurial task. Informal communication describes those communication contacts which are not conditioned by organizational regulations. They influence the social relations between the interlocutors and the atmosphere of the communication. In the case of largely consistent circumstantial conditions this is characterized by a *balance* that is typical for enterprises, which in the case of changing conditions often drifts to the disadvantage of informal communication. Informal communication is frequently one of the first measures taken to respond to altered circumstances within the production process [15.142]. As a result, typical phenomena that appear include:

- Increasing bustle and consultations within production and attached areas
- Accumulation and temporal extension of meetings (for example, production conferences)
- Increase of the error ratio due to an increasing proportion of informal communication

Being carried by the aforementioned established procedures for information processing, information can be subdivided into data, signs, and signals and thus transferred by coded means. The fundamental information process – as a preliminary stage of communication and knowledge development – is executed in three steps:

- Information gathering
- Information processing
- Information transfer

Based on the information process and the information gathered through it, the communication process can be explained according to the model by *Shannon* and *Weaver* (Fig. 15.54) [15.143]. Production and the necessity to organize the corresponding information and communication processes lead to the need to establish appropriate information processes and communicative relations, which should contain as high a fraction of formal communication as possible.

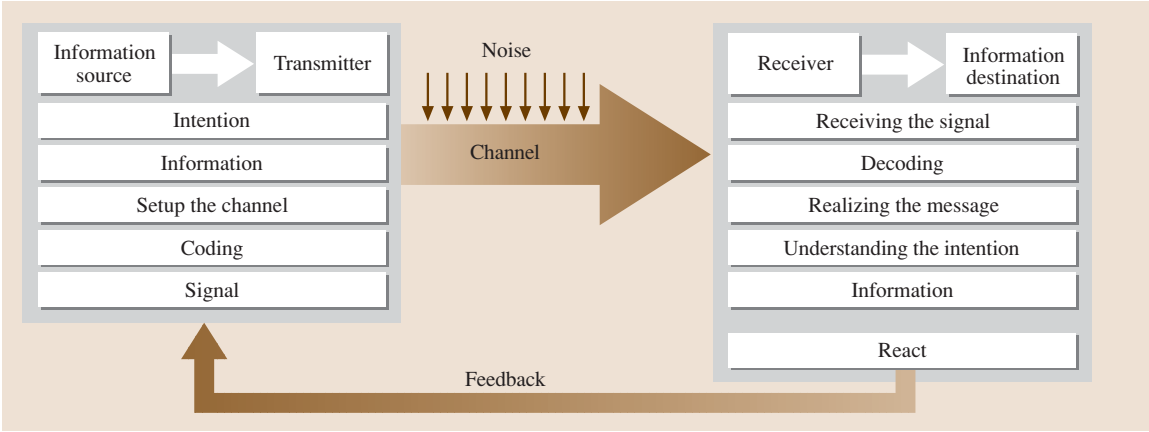


Fig. 15.54 Enhanced model of communication (after [15.143])

15.9.2 Challenges Concerning the Internal Embodiment of Communication Processes

For the embodiment of internal communication processes it is necessary to combine the diverse information demands required for the goal under consideration with the available information. To achieve this, logistic principles should be provided, such as:

- The correct information (related to purpose and target)
- The proper quantity (a random sample or full check)
- The proper quality (i. e., content, significance [processing], measurement errors, error ratio, etc.)
- The correct time (i. e., up-to-date)
- The right location (local availability and possibility to make accessible)

Cost-related considerations as well as the optimal employment of capital must be taken into account in this respect.

During the conceptual design of communication processes, the following specified and thematically grouped problem areas arise:

- Content–conceptual problems
- Organizational–structural problems
- Resistance in personnel and corporate culture [15.144]

These obstructions can be specified more closely and form the basis for the application of a systematic approach when conceptualizing internal communication processes.

The most frequent content–conceptual problem areas are facts such as:

- Fundamental neglect of internal communication
- Lack of coordination of the company's objectives with the means of communication within the company
- Incomplete or even nonexisting description of the target group for the communication (i. e., employees)
- The lack of verification of the result of applied measures as well as insufficient assignment of effort to certain measures

Content–conceptual problems frequently arise from methodically inadequate approaches of installing communicative tools in the operational routine. In addition to these methods being one-dimensional, certain tools are often not harmonized with each other and commu-

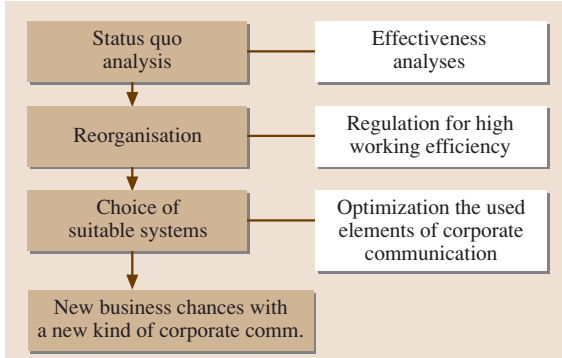


Fig. 15.55 Basic procedure of the corporate communication analysis

nication is not a successive process. An apparent sign of this is the neglect of the corporate identity (CI) concept during the internal communication, for instance. Moreover, further approaches for avoiding problems arise from the existing company structures and organizations described above, which are caused by:

- The lack of institutional and formal rules for voting and decision (such as, tasks without continuance, definite rules as job instructions, obligatory nature)
- The lack of organizational fixation and improper handing over of responsibility when coordinating diverse communication tools (contact point)
- Problems in the case of prompt integration of subsidiary companies with increased independent status (content-related summary)

This is accompanied by obstructions in personnel and corporate culture such as:

- Incomplete understanding of integrated communication in top and middle management
- Existence of subcultures with highly different ways of thinking and behavioral patterns
- Fear of loss of competence
- Lack of feeling for integration
- Lack of professionalism of the responsible personnel
- Lack of willingness to cooperate and coordinate, and to receive information
- Fear of supervision

15.9.3 Methods of Embodiment, Organization Models, and the Management of Communication

The following elements of embodiment are regarded as the result of corporate development in the company and as general limitations to this embodiment:

- Using ambition for reward (striving for gratification, meaning not only money)
- Response to individualization and regional peculiarities (considering individual and regional–cultural behavior patterns)
- Participation and integration of all people concerned (integrating employees into the process of embodiment)
- Focusing on dialogue (mental stimulation, identification, and appreciation)
- Credibility, objectivity, and comprehensibility of both information and the communication process [15.145]

Furthermore, the following assumptions apply to its conception and realization:

- Feedback marks the core of communication.
- Information also means the obligation to acquire it.
- The information portfolio within the company has to be transparent.
- A communicative infrastructure forms the basis for systematic communication.
- The combined employment of media and tools is essential for successful communication.
- Transfer of information is related to the demand and target group.
- Information ought to be clear, intelligible, and restricted to the essential facts.
- Communication complies with an integrative appearance of the company.

Being a decisive factor, the mode of organization of operational procedures has a major impact on the design options for communicative processes within an enterprise. Basically it is the case that, the higher the degree of autonomy of an organizational unit, the higher the demands concerning the availability of information and, accordingly, concerning the required communication process (see also Sect. 15.1) [15.146].

Based on these preliminary remarks, the organization of the structure and procedure of corporate communication gains special importance in securing the realization of the company's objectives on a long-term basis. The corporate vision of the company in conjunction with the strategic goals and the development of resulting essential achievements lead to significant input parameters for the purposeful embodiment of communication processes [15.147].

The need for steady improvement of corporate communication requires frequent changes of structural and procedural organization patterns, since these – as described above – significantly determine the embodiment options, efficiency factors, and effects of communicative media. Generally, the procedure in this respect occurs as presented in Fig. 15.55.

Methods used for these analyses are:

- Portfolio analyses
- Cross-impact-analyses
- Questionnaires, for instance, concerning employee satisfaction

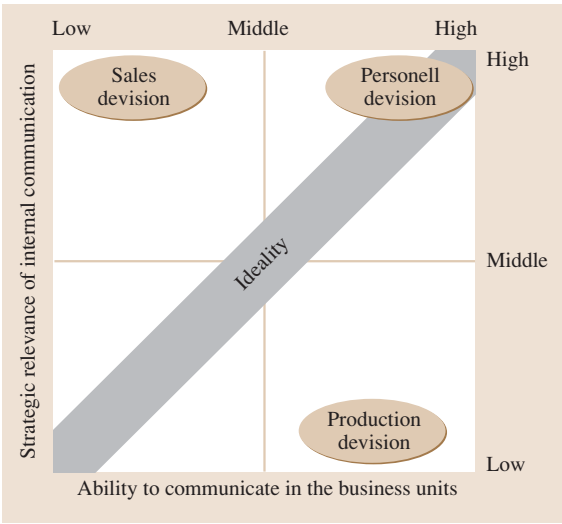


Fig. 15.56 Portifolio analysis for the evaluation of tregetic business units

- Document analyses and analyses of the flow of information
- Checklists concerning the corporate culture
- Analyses of strengths and weaknesses
- Surveys concerning management behavior or methods

Whereas portfolio analyses aid in the alignment and embodiment of business units, cross-impact-analyses are used for the prioritization and consideration of different communication measures for a preassigned goal. Checklists concerning the corporate culture and analyses of strengths and weaknesses confront demand-related rating scales with the rating that had currently been drawn, thus indicating the necessity for development and activity.

Portfolio analyses are used for the observation of strategic business units regarding compliance with communicative demands to fulfil strategic goals. Relations between the strategic significance of a business unit and its determined capability for communication are evaluated by using this subjective method [15.148].

By compiling evaluation items and their *attributes*, a group of experts determines the dimensions of the portfolio's axes. After assessing the strategic business units, the actual value is compared with the desired value (ideality). From the items' deviations from ideality, indices and prioritizations of basic measures can first be extracted (Fig. 15.56) [15.149].

Main result of the portfolio analysis is the extraction of a generic communication strategy to allocate the best improvement measures the company's business units. The success of these measures mainly depends on their probability of occurrence.

Evaluation matrix

Exertion of influence \ Level of impact	Commercials	Promotion	Public relations	Personnel connections	Direct-marketing	Sponsoring	Event-marketing	Exhibitions	Internal communications	AS	Q
Commercials (1)		4	3	2	2	1	0	0	0	12	1,71
Promotion (2)	2		0	2	1	1	1	1	0	8	0,73
Public relations (3)	1	0		0	0	3	3	1	1	9	0, 90
Personal connections (4)	0	2	0		0	1	1	2	1	7	0,54
Direct-marketing (5)	0	0	0	1		0	0	1	0	2	0,40
Sponsoring (6)	3	3	4	2	2		3	2	2	21	1,91
Event-marketing (7)	1	0	2	1	0	3		2	3	12	0,92
Exhibitions (8)	0	2	0	4	0	0	3		0	9	0,90
Internal communication (9)	0	0	1	1	0	2	2	1		7	1,00
PS	7	11	10	13	5	11	13	10	7	87	
P	84	88	90	91	10	231	156	90	49		
AIS	9,66										

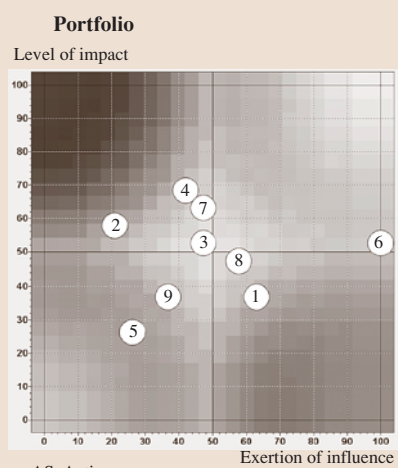


Fig. 15.57 Evaluation matrix and portfolio of the cross-impact-analysis (AS – active sum; PS – passive sum; Q – level of impact (Q = AS/PS); P – exertion of influence (P = AS · PS))

The cross-impact-analysis addressed that problem. By means of a formalized, qualitatively supported method, it enables the evaluation of different incidence rates of states by experts, making it tangible as the embodiment method at the same time. The observed incidents (measures) are evaluated depending on the direction and strength of the context and the diffusion period. Accordingly, the probability of an incident occurring and its impact on another incident can be related.

A team assembled of mainly interdisciplinary experts evaluates:

- The incidents considered to be relevant
- The individual rates of each incident
- The moment of occurrence of each incident in the case of a given incidence rate of 0.5
- The influence of an incident on another

Based on this evaluation, an influence matrix can be constructed, thereby recording the intensities of the incidents with respect to each other (Fig. 15.57). For the evaluation of the influence intensities, a closed discrete rating scale is used.

From the summation of the line or column sums – respectively labeled as active and passive sums – and their relation to the average influence sum (that is AS or PS/number of elements), conclusions can be drawn about the influence probabilities of certain measures. Based on the relation between these figures, the elements can be subdivided into *active* elements, *reactive* elements, *critical* elements, and *inactive* elements, thereby providing clues for the utilization of the measures. Thus, the term of active elements (AS < AIS, PS < AIS) for instance means that the measures have to be changed deliberately, since these determine the entire system in question [15.150]. The elements are characterized by cross-impact-analysis as follows

$$\begin{aligned}
\text{active elements} \quad & \text{AS} > \text{AIS}\,, \ \text{PS} < \text{AIS}\,, \\
\text{reactive elements} \quad & \text{AS} < \text{AIS}\,, \ \text{PS} > \text{AIS}\,, \\
\text{critical elements} \quad & \text{AS} > \text{AIS}\,, \ \text{PS} > \text{AIS}\,, \\
\text{inactive elements} \quad & \text{AS} < \text{AIS}\,, \ \text{PS} < \text{AIS}\,.
\end{aligned}$$

As a result of this analysis, a realization plan can be issued.

For the realization of communication strategies, a multitude of options are currently available. Some example communication media are:

- Worker and information journals
- Postings, blackboard announcements, info systems, wall newspapers
- Internal information services, circulars, newsletters, infomails
- Specific as well as interdisciplinary information sessions
- Press reviews
- Company broadcasting
- Events such as open days and information sessions concerning the company
- Videos, pictorial documentation, compact discs (CDs)
- Annual reports
- Internet presentations, intranet platforms
- Introductory texts
- Trade-fair stands

15.9.4 Conclusions and Outlook

Production systems constantly exchange relevant information with the external environment. As a result, the quality of internal and external communication is of inestimable value for companies. Following the current thrust towards innovation caused by the widespread usage of information technology, information processes as well as the availability of information as the basis for acquiring knowledge have been accelerated. As a consequence, conceptual and creative solutions are becoming increasingly important, purposefully integrating the new options for the embodiment of communication processes within organizational structures (Fig. 15.58).

Much attention is being paid to interorganizational learning. Through *learning alliances*, firms can speed up capability development and minimize exposure to technological uncertainties by acquiring and exploiting knowledge developed by others. However, learning introduces self-organization, which makes it difficult to guide.

Modeling agents [15.151] help discover such emergent behavior inherent to a multi-agent situation consisting of large groups of loosely coupled agents that work together on tasks. Interoperability of agents is required, and their negotiation strategy will depend on their organizational role (motivational quantity) [15.152].

Collaboration may be expanded beyond the current concepts into complex organizations spontaneously emerging from dynamic versatile environments [15.153,154]. Self-organization and self-similarity help to reduce organizational and process complexity by fractalization. By using topological and systemic methods, supported by the theory of complexity, power-

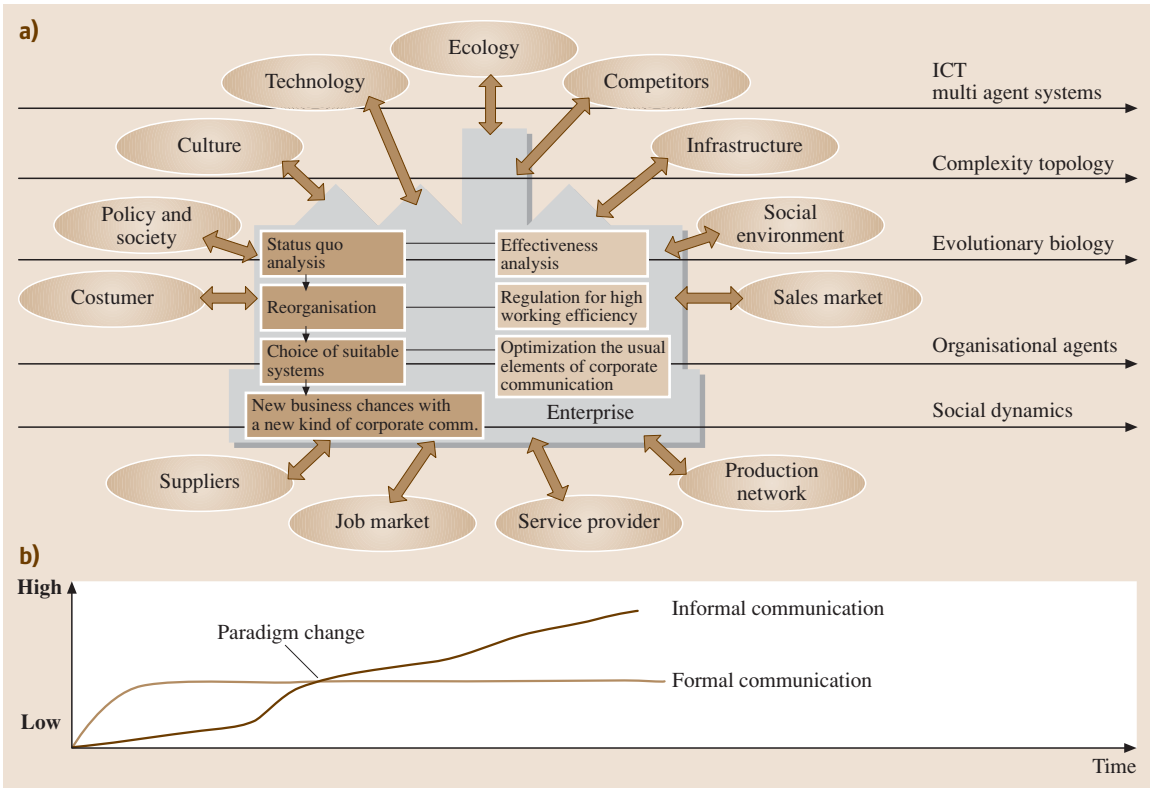


Fig. 15.58a,b Communication impact and research strains in organization and process engineering

ful organizational production network optimization and control conjectures may be provided. The steady need for improvement of corporate communication assumes versatility of structural and procedural organization patterns [15.155].

Self-organization is also a specific topic in social dynamics between organizational entities envisioned as populations of interacting agents in a given structure [15.156]. At the level of departments or even whole firms unofficial structures may emerge from the social interactions of individuals. Collaborative teams may increasingly turn the enterprise into densely interconnected networks whose evolution is driven by simple universal laws that may resemble evolutionary patterns [15.157]. In this sense, collaboration is envisioned as a strategy for the fitness selection of an entity.

Relations between the strategic significance of a business unit and its determined capability of communication are evaluated by impact analysis [15.150]. Procedures for corporate communication and negotiation gain special importance in securing the realization of enterprises objectives on a long-term basis [15.158]. The corporate vision of the company in conjunction with its strategic goals lead to significant input parameters for the purposeful operation of strategic networks [15.147].

The huge field of enterprise organization and operations is anything but static and is composed of numerous threads of research. Technology is changing,0 therefore significant changes in organizations are most likely in the future. Ultimately information and communication technology seems to be the key driver determining how and when future developments will occur.

15.10 Enterprise Collaboration and Logistics

15.10.1 Dimensions of Enterprise Networks

Industrial collaborations occur because of economic necessity. Modern manufacturing imperatives are the production of complex goods in a highly competitive market environment where cost, quality, and time to market are *facts of life*. In this, so-called buyers', market of sophisticated goods a single enterprise can hardly successfully compete without collaborating with other enterprises that complement its competencies, goods, and/or services. To meet these challenges, companies have learned to focus on their key competence features (the *focused factory*) and collaborate with complementing enterprises to produce a complete range of products and services to meet ever-changing market needs.

The increased dynamic of changes demands quicker orientation in dynamic, unstable markets. To achieve economies of scale global markets have to be addressed. Products and production processes have to be adopted according to the emerging diversity of these new markets. Based on the increasing complexity and heterogeneity of products the critical size of a single enterprise for autonomous engagement with the market has changed dramatically. This problem can be solved by appropriate industrial cooperations. Especially for small and medium-sized enterprises (SMEs) or even medium-sized companies, cooperation with other enterprises, whether they are competitors or not, often seems to be the most promising approach for success in future markets. As the impacts, motivations, and requirements for industrial cooperations are subject to alterations, the variety of real-life applications will increase dramatically in the future and will become more difficult to anticipate.

Industrial collaborations are not new; they have been well known for many years, even decades. In the one-of-a-kind business, product- or project-specific industrial cooperations were introduced many years ago. One-of-a-kind-production (OKP) companies very often see themselves more as system integrators for customer-specific products and less often as manufacturing companies dealing with the whole range of industrial manufacturing including all parts and component manufacturing activities. This has been recognized as a critical asset for serial and mass production only in the last 5 years. Accordingly many approaches as well as methods have been known for years, with their roots in the one-off domain. Well-known examples are the formation of temporarily limited enterprises in the large-scale engineering, construction, and shipbuilding industries.

However, the relevance, need for, and chances of success of industrial cooperations have changed dramatically and will become increasingly important in today's information society.

Industrial cooperations are a complex multidimensional phenomenon. As there are finally no restrictions at either political, economical or technological level that could prevent certain types of industrial cooperations, the potential variety is almost unlimited. Accordingly an unambiguous and clear distinction or classification of typical types of cooperations will become difficult or even impossible. However, the reduction of today's vast variety of cases to a small number of types is not sufficient to gain a better understanding and systematic overview of the phenomenon of industrial cooperations.

Analyzing typological issues of enterprise networks will support:

- Systematic problem analysis and solution synthesis with respect to cooperating enterprises
- A more structured view of the vast amount of possible real-life cases of industrial cooperations through systematic characterization of these cases
- The analysis of cases and relevant types of industrial cooperations with respect to typical problems related to decision-making, production planning, etc.

Nature of Cooperation

In general to cooperate means to act or to work together. Cooperation between companies materializes for one purpose and one purpose alone: self-interest. Potential motivations for setting up a network of cooperating enterprises are short-term opportunities as well as an intended optimization of cost, time, and/or quality. Accordingly the added value of a cooperation must be measurable for the cooperating companies. Cooperation will usually involve the exchange/sharing of goods or services or any imaginable combination thereof. According to the literature, there is no common understanding about the term *cooperation*. However, with

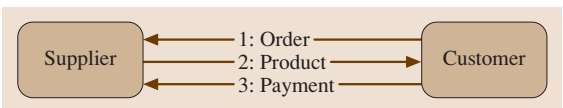


Fig. 15.59 Steps of a fundamental market transaction

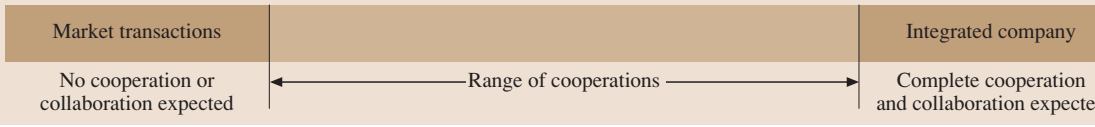


Fig. 15.60 Range of cooperations

respect to production management, two major specifications can be distinguished:

1. Cooperation in a broad sense comprises all types of collaboration between organizations that participate in the economic life.
2. Cooperation in a narrow sense characterizes the close interorganizational collaboration between autonomous organizations.

One of the fundamental market transactions is between two parties that agree to conduct a one-off market transaction. This transaction (Fig. 15.59) consists of the customer placing the order, the supplier supplying the product/goods, and finally the customer reimbursing the supplier for the goods received. Of course, this three-stage transaction could be conducted in one single step as transpires when an individual makes a purchase in a shop, or, based on the credit worthiness (as judged by the supplier) of the customer, during an agreed time scale. For the purposes of our discussion, the issue here is not the time scale or whether the transaction is conducted face-to-face, but the freedom to conduct the business once this transaction is complete. In the fundamental market transaction, the customer is free to seek another supplier for the following transactions for the same products/goods.

The concept of a contract in this scenario is fairly straightforward. When goods and payment are exchanged it is implied that the goods conform to some *expected standards*. These standards could be mutually understood or defined by the supplier or by legislation (consumer protection); for example, if the product is a software program, it is mutually understood that the supplied program will work according to the given specifications and that it will not contain any bugs.

Similarly, claims made by a supplier for its products in advertising form an implicit part of this transaction. Therefore, if the goods exchanged are foodstuffs, legislation dictates that such goods are of edible quality. The term *expected standards* is never fixed, since there is no explicit written contract. Hence, based on the perspective of the supplier and customer, it could be disputable. In those cases, one normally needs the services of an intermediary party or eventually the courts to ascertain the facts and resolve the dispute.

Aiming at a detailed systematic specification of types, attributes, dimensions, etc. of industrial cooperations, the term cooperation will be understood as an alternative type of transaction, bounded by the transaction types *market* on the one hand and *integrated company* on the other hand (Fig. 15.60). These two types of cooperations, market and integrated company, depicted in Fig. 15.60 are the idealized extremes. All possible types of cooperations that can materialize between any two enterprises lie within these two extremes.

At the market level only single market transactions occur. The parties have no obligation to repeat the transaction or continue the business. However, if they continue to conduct transaction-based business over time, confidence between the parties could increase, re-

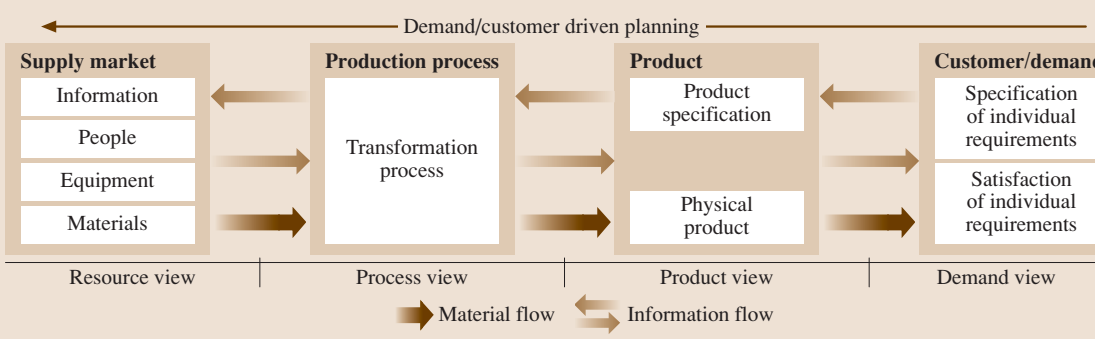


Fig. 15.61 Information and material flow in a demand/customer-driven manufacturing system [15.159]

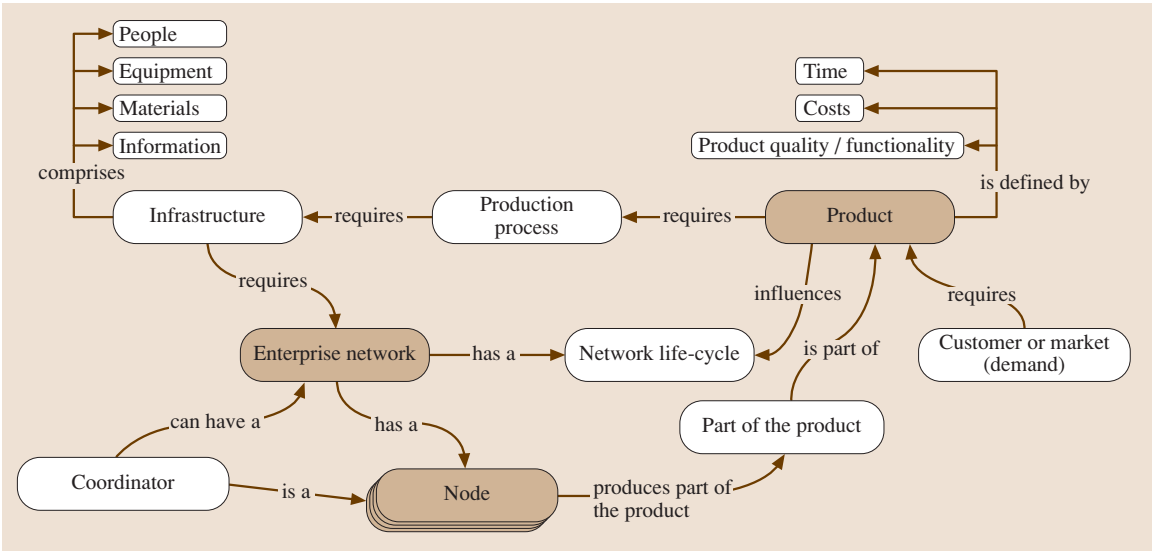


Fig. 15.62 Conceptual model of a generalized enterprise network

sulting in the breakage of the explicit link between the exchange of goods and the payment; for example, once the credit worthiness of the customer is established, the invoicing and payment cycles need not synchronize with the dispatch of goods. On the other extreme, all parties belonging to a single enterprise are expected to cooperate fully and direct their efforts towards a single goal.

Conceptual Model of Enterprise Networks

Figure 15.61 depicts the interrelations between customer (demand), product, production process and manufacturing resources from a system point of view.

In general, products can be described by three significant attributes: quality and product functionality, time (availability on market), and cost (price on market). The underlying demand for a product can occur in two ways: A customer requires the product and comes to the market for its purchase. Secondly, a market survey by a manufacturer ascertains a need (a potential market) for a product, which triggers its production. Since the trend is towards customer-driven manufacturing, these three attributes will be influenced directly or indirectly, by the end (or perceived) customer. Once the attributes and saleability of a product are established, a production process is required for its manufacture, which will in turn require manufacturing infrastructures to manufacture it to the specified attributes.

For decades the classical approach of an enterprise was to invest in the required resources and thus to realize production process using its own resources. This has changed dramatically. Today no single enterprise is able to provide all manufacturing resources as well as competencies necessary for the realization of ever-changing customer demands. Enterprise networks as a dynamic, interenterprise configuration of manufacturing resources and competencies have become a promising alternative to offer the required manufacturing infrastructure. In the following the term *enterprise network* characterizes a suitable collaboration of two or more cooperating enterprises, aiming at a common and collaborative realization of a certain product and/or service. Thus, the term enterprise network is always related to a specific product and/or service.

Different partners of the enterprise network take responsibility for the different elements of the value chain. Therefore, the generation of added value can take place simultaneously and at various locations. In an interlinked manufacturing environment, customer and supplier relationships becomes critical assets. Knowledge about the capabilities of potential partners is becoming a part of the knowhow of an enterprise.

Based on the dependencies within a customer-driven manufacturing system, as depicted in Fig. 15.61, a conceptual model of a generalized enterprise network is documented in Fig. 15.62. Elements highlighted in the figure are discussed in detail in the following sections.

Each enterprise network will have distinct and clearly defined set of members (nodes), which are

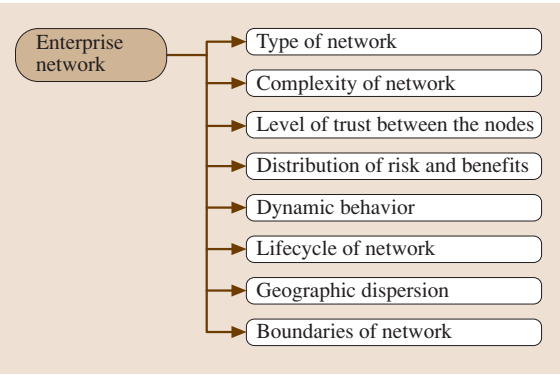


Fig. 15.63 Enterprise network-related attributes

usually independent enterprises. Once a network of enterprises decides to cooperate, this network will have a distinct lifecycle, which will be influenced not only by the corresponding priorities of the cooperating nodes but also by the product–market combination. Furthermore, each node will have its own objectives, which will influence the primary product–market combination that brought this node into play in the first place. The function of the coordinator is to manage the dynamics of the enterprise network. Quiet often the role of coordinator of the whole enterprise network is taken over by the final assembler. However, there may be circumstances in which an external specialist agent can perform this role more effectively.

The goal of this model is to act as a baseline for a systematic view of enterprise networks and to give a systematic structure to the vast number of attributes to be considered while forming a new or analyzing an existing enterprise network. However, it is not intended

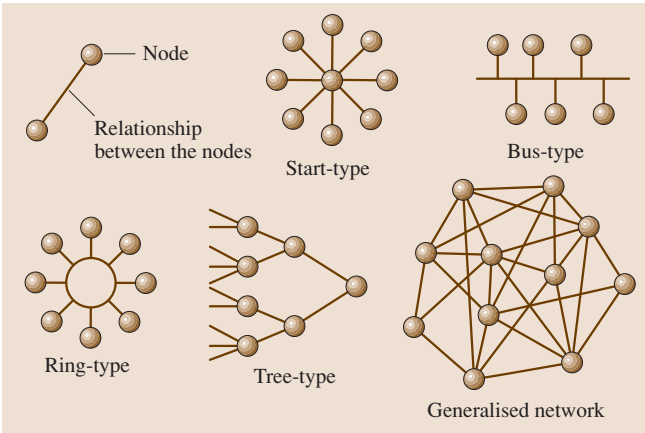


Fig. 15.64 Basic enterprise network types

that the conceptual model be used as a baseline for an overall model including all possible aspects of enterprise networks.

Enterprise Network-Related Attributes

Aspects that characterize the overall network from a systems perspective are considered under the category of *enterprise-related attributes*. This discussion is restricted to the attributes given in Fig. 15.63.

Type of Network. In general, networks are defined by nodes and relationships between these nodes. As already stated the minimum amount of nodes in a network is two. By considering more than two nodes various types of enterprise networks are possible. Figure 15.64 depicts some well-known types of networks.

The communication between any two peripheral nodes in a star-type network will always be conducted through the central node. Therefore the central node could be considered as a *controlling* node of a star-type enterprise network. A bus-type enterprise network implies some form of flow of goods or information from left to right. In ring-type enterprise networks there is no unique direction of the flow of information or products, which instead can take any path. Furthermore, the feature distinguishing the ring from the star type of enterprise networks is the absence of the central controlling node. Therefore, in ring-type enterprise networks, all nodes are hierarchically equal and any two can communicate directly. Tree-type enterprise networks can be converging (as shown in Fig. 15.64) or diverging (a mirror image of the converging type). In either case of tree-type enterprise networks, the flow of information and goods is usually assumed to be from left to right. The controlling node in a converging tree-type enterprise network is often the one downstream to the operations, with overall responsibility taken by the extreme right-hand node. Diverging tree-type networks are often distribution-type networks. In this case the controlling node if more often than not the extreme left-hand node. A generalized network is a complex interrelationship among several nodes. The connections between the nodes and the issues of controlling node cannot be generalized and predefined, and they are situation and case dependent.

Complexity of a Network. In a broad sense the complexity of a network, as a multidimensional attribute, can be characterized by the number of nodes, the number and types of the relationships between these nodes, the dynamic behavior of the relationships, and the num-

ber and types of products and services provided by the network. By its very definition, there is no physical limit to the technical complexity of the network. However the number of relationships might be limited by the technology to be applied, and the dynamic behavior within the network might limit the optimal number of nodes in a network. The greater the number of partners, the more difficult it will be to manage. Usually, in such complex cooperative networks, the node downstream (customer) will initiate or chase the node upstream (supplier) for goods/services. This request may be triggered internally from within the customer's operations or externally by a third party.

Bilateral relationships are used as elementary building blocks to create complex cooperative enterprise networks. Real-life enterprise networks can be seen as more or less complex combinations of various types of bilateral relationships. In general, networks may be composed of equal or different types of bilateral relationships (homogeneous versus heterogeneous networks). To reduce the various types of interorganizational relationships known from the literature according to *Jagdev* et al. five basic types are distinguished in the following [15.1]. Accordingly enterprises delivering/supplying parts or components based on a market transaction will not be considered as being a member (or a node) of an addressed enterprise network.

Figure 15.65 depicts the decompositions of a large heterogeneous network (A). Network A is decomposed into a series of subnetworks within the defined terminologies of SE (a single enterprise located at two different geographical locations), VE (network D), EE (network C), SC (network B), and MT (network A).

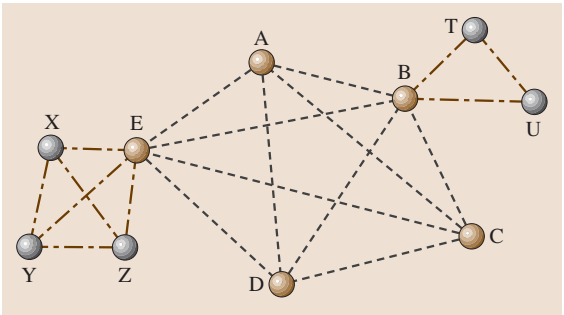


Fig. 15.66 Direct and indirect relationships of the nodes of a network

The complex dependencies within complex networks or between networks are illustrated in Fig. 15.66. Although there is no direct link between node A and nodes U and T, via node B both nodes might be important for the success of the network in which A (and B) is involved.

Level of Trust Between the Nodes of a Network. Trust between the cooperating partners is a prerequisite for a successful partnership within an enterprise network. Usually, when two parties start a partnership (i. e., agree to do business together), they start with some basic level of trust as expected by the norms of the business environment. However, as the partnership flourishes, trust builds and forms a foundation for an

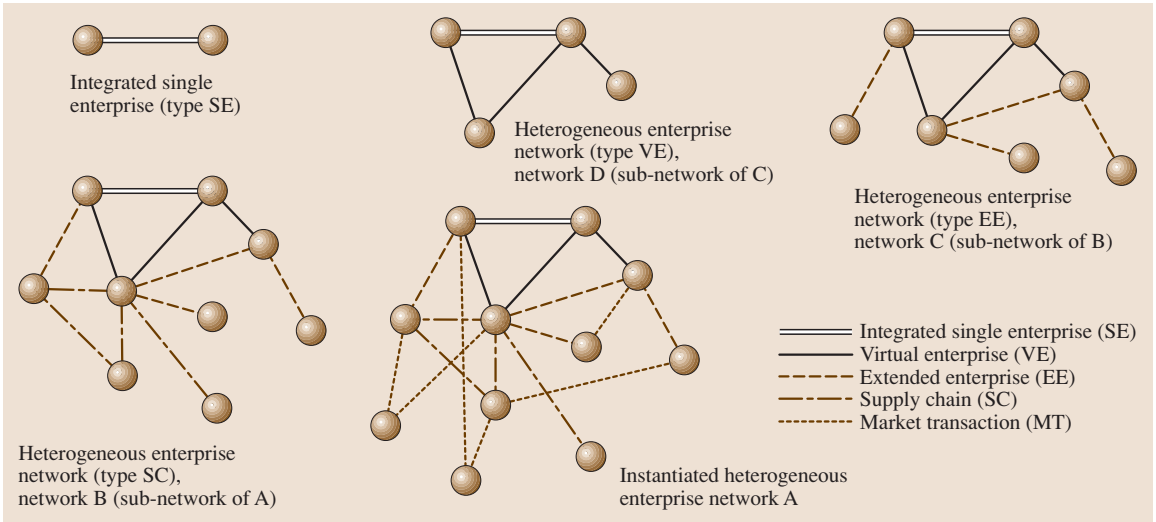


Fig. 15.65 Describing the complexity of enterprise networks

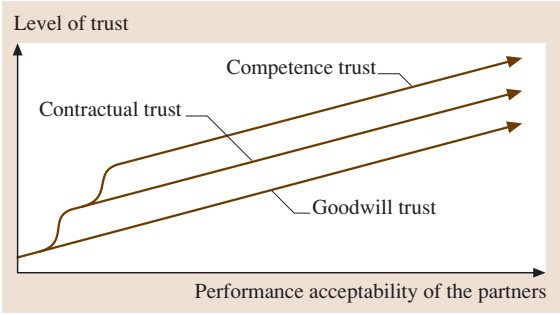


Fig. 15.67 Levels of trust

ever-closer and mutually dependent business relationship. Apart from the *basic trust* expected by business ethos (and presumably understood by the parties involved), the level of trust – which can have, for all practical purposes, no limit – is closely linked to the behavior of one party towards the other. Real trust builds under extraordinary circumstances in which one partner is willing to meet exceptional requests, above and beyond the agreed terms of business, so that the other partner is not let down. It is such particularly testing situations that establish the real level of trust between the partners. For example, partner A has agreed to supply partner B with an agreed amount of goods per week. Because of exceptional circumstances (such as a new important customer for partner B), partner B suddenly requires way above the agreed amount of goods for the following few weeks. If partner A is willing to go to extra lengths (say, by increasing overtime levels) to meet partner B's increased requirements, then partner B will perceive partner A to be more reliable and hence more trustworthy.

These subtle behavioral responses of partners play an essential part in the development of trust. This is particularly so in the development of perception of trust. Authors believe that perceived trust is often more important that trust itself. It is the perception of trust – i.e., how one party perceives (never mind the reality) the trustworthiness of the other party – that determines the evolution of longer-term and ever-closer business partnerships.

Childe [15.160] classifies trust into three headings: goodwill, contractual, and competence trust. *Goodwill trust* is that by which a partner is trusted to take decisions without unfairly exploiting the other partner. *Contractual trust* is the keeping of promises, such as delivering goods or making payments on time, or maintaining confidentiality. *Competence trust* depends upon the technical and managerial competence of the company to perform a function, such as delivering components within specifications.

Normally the development of various types of trust follows a clearly defined trajectory, as depicted in Fig. 15.67. Companies agree to do business based on some level of goodwill trust. As the partnership progresses, the performance of each partner becomes validated by the other and, if acceptable to both parties, they agree to formalize their collaboration in the form of longer-term contracts. As the contractual agreements progress, competence of the partners comes into play, together with the resulting evolution of competence trust. Therefore, a high level of competence trust would need to exist for a supplier to be allowed to make deliveries direct to the assembly line with no inspection.

Distribution of Risks and Benefits. As for a chain, a network is only as strong as its weakest node. Therefore, to optimize the performance of the whole enterprise network, one not only has to optimize individual nodes but also to optimize the network as a whole. Hence, a balanced distribution of risks as well as profits within the enterprise network is a prerequisite for successful cooperation. As an enterprise network can deliver a broad range of products and/or services the distribution of added value achieved within a network might vary from order to order. Examples are the increasing number of networks of SMEs, craft organizations, etc. offering full services in the area of building construction and maintenance of energy/water supply systems.

The survival of a network is endangered if there is no added value to be achieved by the overall activities of the network. In general partners who cannot perceive any added value by cooperating with others will concentrate on their individual activities and will not put sufficient effort into the network. Mutual benefits in terms of shared added value become a stabilizing factor for cooperative behavior in a network [15.161]. Accordingly, by supporting improvement at the supplier rather than trying to drive down the price, the principal company in a supply chain assists in reducing costs rather than reducing the profit at the supplier [15.160].

Dynamic Behavior of a Network. Bilateral relationships as well as the network itself will often change over time according to the maturity of a product, the stability of a market, the priorities of the partners, etc. Depending on their mutual interests, two independent companies could merge into a single enterprise. On the other hand, market conditions may compel a large company to sell of its noncore divisions. Between these two

extremes an almost infinite variety of *getting together* and *going apart* scenarios is possible. A network need not be stable over time; indeed it rarely is. The functions of a node may evolve over time, which may influence its place/role in the network. Even though the scope collaboration between the nodes of a network may be very large, the relationship can be categorized within well-defined paradigms, such as supply chain, extended enterprise, and virtual enterprise. Section 15.7 elaborates further on these categories.

Lifecycle of an Enterprise Network. Just as a product has to pass through the various stages of its lifecycle (from idea to recycling) the lifetime of an enterprise network also depends on the success of the offered good or service and can be described by at least four significant lifecycle phases (Fig. 15.68):

1. Preparation (sourcing of partners)
2. Setting up (legal issues, contracts, etc.)
3. Operation (day-to-day management of the network)
4. Decomposition

A prerequisite for frictionless operation is efficient interorganizational product data and process management [15.162]. The duration of the operational phase of an enterprise network can be classified based on the premise and understanding under which the nodes of the network cooperate.

Formal Duration

The formal duration of a cooperation describes its contractually fixed duration. It can be classified as unique if the intention is to realize just one product/offering based on a specific customer request. The cooperation can be classified as limited if the intention is to realize a fixed series of products or a product line, or it can be unlimited. Accordingly groups of enterprises might cooperate in a network environment with the aim of offering an additional product or service to the market. An enterprise network can be seen as a problem-oriented configuration of competencies according to market/customers needs [15.163]. Based on an actual request, some nodes of the network will fulfil the order along on order-specific value chain.

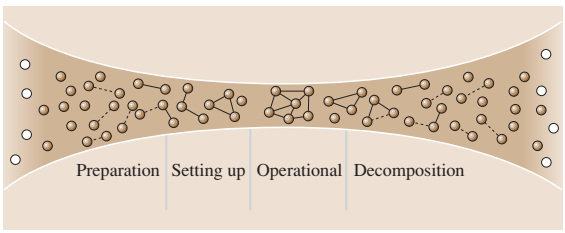


Fig. 15.68 Significant lifecycle phases of an enterprise network

Traditionally the duration of a cooperation between companies is linked to the execution of a customer-specific order and the delivery of the related product. As, especially in the capital goods domain, problem solutions instead of *pure* products tend to be requested, product-related responsibilities are extended into the operational phase and the disassembly of the product as well. Relationships of cooperating enterprises and the related cooperation may change over time, either improving or deteriorating. In other cases, a cooperation might be the preliminary stage for an integrated organization (e.g., the takeover of a cooperating partner). Nevertheless, one can expect that the variety of cases will still increase whereas the formal stability and the formal duration will decrease. However, if the formal duration of a cooperation is coming to an end the dissolution of the enterprise network should be clean, amicable, and with the minimum of damage to either partner.

Informal Duration

The informal duration of a cooperation is characterized by the mutual confidence that enterprises put into the cooperation. Whereas a formal cooperation is actually restricted to the delivery of one product, the informal agreed cooperation might be unlimited. In this context it should be mentioned that, from the game-theoretical point of view, the repetition of an interaction or even the expectation of a repetition is essential for someone's cooperative behavior. People act in a cooperative manner if they expect an additional interaction; if not, noncooperative behavior might be more effective and more successful [15.164]. Taking learning effects into account, both the first and second cooperation (even if it is a *second unique* one) will make a profit from a cooperation based on mutual confidence. The informal duration of a cooperation might be supported by contractual agreements such as cooperation agreements, letters of intent, etc.

15.10.2 Analysis of Enterprise Collaborations

Bilateral Relationships and Enterprise Networks

When two or more enterprises collaborate, they could form an enterprise network. Any enterprise network, however complex, is composed of a series of bilateral

relationships. Each of these bilateral relationships can be either hierarchical or nonhierarchical (Fig. 15.69). The relationship is hierarchical in nature when one party (who could be a buyer or a seller, depending on the market environment) is clearly superior to the other and makes the rules of engagement. In a nonhierarchical partnership, the two parties are equal in status, operate as partners (such as those engaged in codevelopment), and neither of them dictates the other. In this case, all decisions affecting the partnership are mutually agreed on.

The relationship between the parties is never static (Fig. 15.69, middle column). Quite often (though not always) partner B would like the relationship to be nonhierarchical in order to have more control over its own destiny. Therefore, it will try to improve its negotiating position, say by offering improved solutions or services. The eventual terms of engagement will be determined by the market conditions of the product or service in question and any *trump cards* the parties may hold. Reputation, knowledge, and technology within the organization, the type and uniqueness of products, and brand names and patents held by the enterprise are some of the examples of trump cards by which an enterprise can leverage its negotiating position to its advantage.

Individual enterprises are therefore required to extend their resources, their control structures, and their information systems, while protecting their market niche, in such a way that they become an attractive partner for organizations that offer products that are complementary to their own. They therefore need criteria that allow them to decide with whom they need to cooperate, and in which means and structures they need to invest. In so doing, they are configuring their own network.

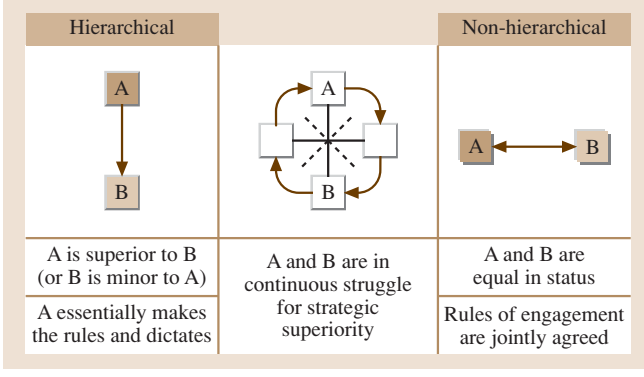


Fig. 15.69 Elementary types of bilateral relationship

Evans and *Wurster* [15.165] illustrate the complexity of interenterprise interactions and interdependencies in their *hyperarchy* model (Fig. 15.70). In a hierarchy each enterprise depends on one superior enterprise, which has access to information that is, by definition, not available to its subordinates. Here the term *hierarchy* refers to the structure of an enterprise network as opposed to a structure within an enterprise. In a hyperarchy type of enterprise network, the communication links can be complex and can take any direction.

Since the bargaining power in buyer–supplier relationships strongly depends on this asymmetry of information, such relationships will drastically change if information technology (IT) eliminates this asymmetry. Evans and Wurster claim that, under the influence of information technology and standardization of communication, hyperarchies will challenge, and eventually replace, hierarchies.

Interorganizational relationships in hierarchies are, by definition, bilateral. When considering networks, thinking in hierarchies therefore typically leads to a strict distinction between customers and suppliers. Customer-oriented effort (i. e., marketing and sales) is directed at ensuring selecting, whereas supplier-oriented effort (i. e., purchasing and procurement) is directed at selecting the right suppliers. However, selecting the right customers can be equally important as being selected by the right suppliers. The hyperarchy model consequently seems a much more accurate representation of interorganizational interdependencies.

The hyperarchy model can be compared to the network models of the International Marketing and Purchasing (IMP) group, which has published widely on networks of organizations cooperating in partnerships [15.166–168]. The IMP group views interorganizational networks as constellations of relationships. Its viewpoint is that each relationship between companies is heavily influenced by the relationships the companies have with other companies. In this view networks are infinitively large, never balanced, never optimal, and have unique perspectives for all their members. The IMP group is not so much studying individual networks, but rather sees networks as a way of looking at the world.

Networks in the IMP view consist of actors and relationships. Relationships are the result of enactments carried out by the actors and can be characterized by factors such as technology, knowledge, social relations, administrative routines, and legal ties [15.167, 169]. Actors are indirectly connected to other actors through relationships. This is called *connectivity*.

Due to connectivity, an actor can only partly identify the networks in which they operates. There are always indirect relationships to other actors, of which the actor is not aware. These unknown relationships can however be essential to the functioning of the actor in question. The IMP group criticizes traditional marketing and purchasing research that primarily focuses on bilateral and direct relationships, whereas they make it clear that multilateral and indirect relationships should also be considered.

However, despite this connectivity, no organization should try to control all relevant relationships, or even be aware of them. It should make a selection of the ones it considers most important and should direct its investments in terms of resources, control structures, and information systems towards these. In order to be able to make this selection, it should be able to identify the various relationships it has, and the alternative forms of cooperation it can use to maintain each relationship. The following subsections describe a typology that helps to characterize and recognize various modes of enterprise collaborations.

Continuum of Enterprise Collaborations

The possible scope of collaboration among enterprises is infinite. There is a plethora of literature on the classification of relationships [15.166,167,170–175]. Many of these classifications are based on the transaction-cost theory, in which transaction costs are weighed against coordination costs. One possible way of locating any collaborative relationship on a continuum is depicted in Fig. 15.71. The motivation behind this classification is to isolate those types of collaborations that can be considered as a part of the enterprise network. By enterprise network we mean *two or more participating enterprises that are engaged in the supply and receipt of goods or services on a regular and ongoing basis, with partners relying on each other, and between which the supply of goods (or services) is constrained by the associated logistics, manufacturing commitments, and operating dynamics of the participating enterprises.*

In the case of market transaction-based operations, the parties involved have one goal: maximizing their own profits from the transaction in question. Basically, if they do not consider the transaction sufficiently profitable the only measure they can take is not finishing the transaction and finding alternative sources. At the other end of the continuum, intra-organizational cooperation may directly influence the continuity of the organization. Their goal is to remain in business; their dependence on the intra-organizational structures gives them very little room for manoeuvre. From market to integrated company, the strength of interorganizational bonds increases. Consequently, organizations have to put more effort into setting up and maintaining such a relationship. Characterizations of the positions on the continuum may therefore serve as indicators of the sophistication of interorganizational relationships. The strength of interorganizational bonds is often analogous to the duration of the collaborative relationship.

Market Transaction. A market transaction occurs between two parties, and the nature of the relationship between the participants is strictly transaction based; there need not be any continuity in cooperation. Even for a market transaction there can be a range of collabo-

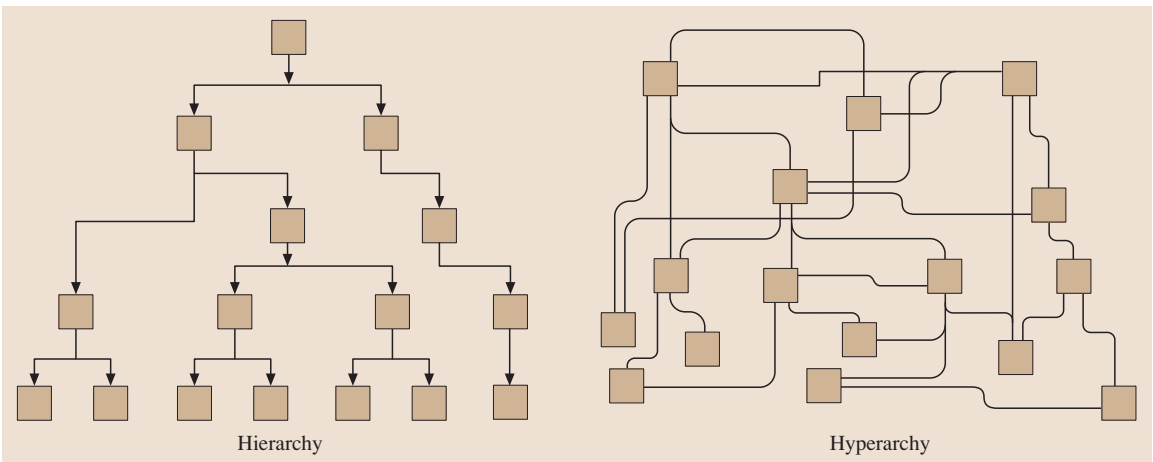


Fig. 15.70 Hierarchy versus hyperarchy

rations; for example, a market transaction could entail a range of collaborations based on the credit worthiness and payment schemes (ranging from payment in advance to mutually negotiated overdrafts and/or discounts) agreed between the parties.

Noncontractual Agreements. Agreements based on trust frequently exist in real life. In a network of enterprises, noncontractual agreements are essentially longer-lasting market transactions. They exist because the customer is happy with the supplier (in terms of pricing, delivery commitments, after-sales service, etc.) and finds no reason to search for alternative sources. Even though there is no formal commitment, favorable prices could be negotiated. The products in such transactions are either standard and readily available in the marketplace, or are not part of the core operations of the customer, for example, a hospital purchasing all of its stationery from one source. For products that are part of the core operations, noncontractual agreements normally evolve into contractual agreements.

A group of companies engaged in noncontractual agreements in the same end-product sector will form a cartel. Cartels are often formed to avoid competition and to maintain a noncompetitive marketing environment. Terms such as *restrictive practices* and *price fixing* are often associated with cartels. Cartels rely on informal understandings that share strategies for mutual benefit; they rarely, if ever, share resources. The relationship between the cartel partners is equal (as opposed to hierarchical) and they are not deemed to be part of enterprise networks.

Contractual Agreements. Contractual agreements between the manufacturers of complementary products and services are very common. Once these agreements materialize the partners will become part of an enterprise network. In a competitive environment formal contracts between the companies engaged in the same end-product sector and competing in the same market should not exist. Some agreements that come close in this category include airline alliances of code sharing, thus *virtually* extending the number of destinations available to prospective passengers of each partner airline. This scenario is only attractive to partner airlines if there is only marginal intersection of the corresponding networks and the union of the new enlarged network complements their existing services, otherwise monopoly and antitrust issues will apply. There can also be a supplier–buyer type of relationship between the producers of similar product, such as BMW supplying engines to Rolls Royce. This again is possible because the customer base for these two car manufacturers is distinct. This type of arrangement, if and when it occurs, will become a part of an enterprise network.

Joint Ventures. A joint venture implies a group of companies supplying complementary services that join forces for mutual benefit. Joint ventures frequently lead to new enterprises with joint ownership, for example, Mercedes Benz and Swatch pooling their respectively unique skills to develop the *Smart Car*. Similarly, Ford and Volkswagen set up a joint venture in Portugal, called Auto Europa, to produce their Sharan and Galaxy vehicles. Such joint ventures normally result in new enterprises that could very well form their own enterprise networks with noncontractual and contractual agreements.

Integrated Company. Due to the fact that, by definition, a single company is supposed to have a single shared ownership, in this case there is unlimited duration of cooperation. The relationship among various hierarchies and divisions of a single company will always take the form of an enterprise network. Even in a single enterprise there can be different levels of integration and cooperation between different subsidiaries or indeed between different departments located within the *four walls* of the enterprise. Although one expects complete cooperative confidence between different de-

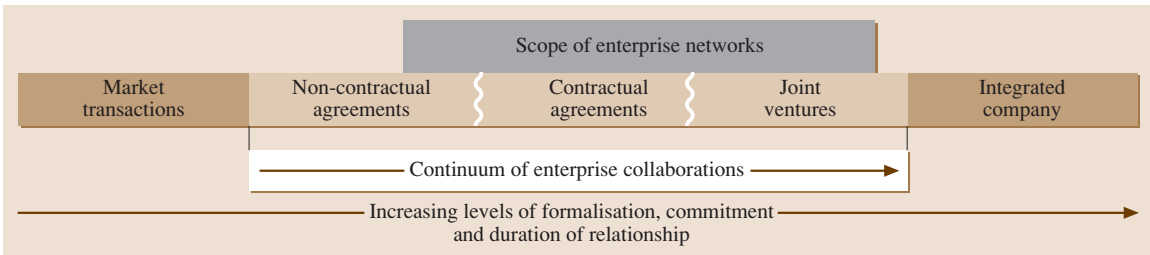


Fig. 15.71 Continuum of enterprise collaborations

partments of a single enterprise, in reality this may not always be the case. Individual aspirations and personalities will play an important part in the building (or hindering) of cooperation.

The types of collaborations described above are neither complete nor exhaustive. The only function of these descriptions is to identify those types of collaborations that can contribute to the formation of enterprise networks; for example, licence agreements could be considered as a form of collaboration, however they are left out of the classification above because they occur between parties engaged in the same product sector or frequently the same product. Partners in licence agreements are rarely part of the same enterprise network and barely share resources. A classical example is that of franchising. Franchising frequently focuses on the *brand name*. In this scenario (for example the fast-food sector), the brand-name owner allows a third party to manufacture/supply products/services under that brand name. In this agreement the brand-name owner will dictate the quality and price of the end-product/service. The relationship in licence agreements is almost always hierarchical.

Types of Collaborations within Enterprise Networks

Formation of enterprise networks is principally dictated by the rule of *return on investment*. This involves the decision of making or buying (outsourcing) some of the components required to complete the product. Two investment areas that can influence this are to avoid the costs of investing in one's own skills (or competencies), and to guarantee the availability of specific products/services. Investment in skills will have to take into account the investment in suitably skilled personnel as well as the associated investment to carry out necessary research and development to keep abreast with the cutting edge of technological developments – indeed push the frontiers further to gain technological advantage. Investment in production resources will be influenced by total in-house production requirements and thus whether economies of scale can be realized. In either case, one way or the other this boils down to the question of whether there is enough internal demand to justify investment in either (or both) of these two areas. Another factor that could influence outsourcing is the issue of time scales: under certain circumstances one could outsource the supply of certain parts, even though they may have the capabilities of making it in-house, simply because an outside supplier can do the job quickly. Whatever the reasons, once the decisions regarding the make, buy or outsource have been made and the supplier selected, this partnership will be a part of the enterprise network.

In this section, the focus of our attention is on the noncontractual, contractual or joint-venture type of collaborations among independent enterprises pooling their core competencies to form enterprise networks. Within the continuum of enterprise networks (Fig. 15.72), three key types of collaborations that can form a component of an enterprise network can be identified:

1. Supply chain (SC) type of collaboration
2. Extended enterprise (EE) type of collaboration
3. Virtual enterprise (VE) type of collaboration

Because there can be an almost infinite number of possibilities according to which companies can collaborate and form enterprise networks, one cannot discretize the space between market transaction and hierarchical

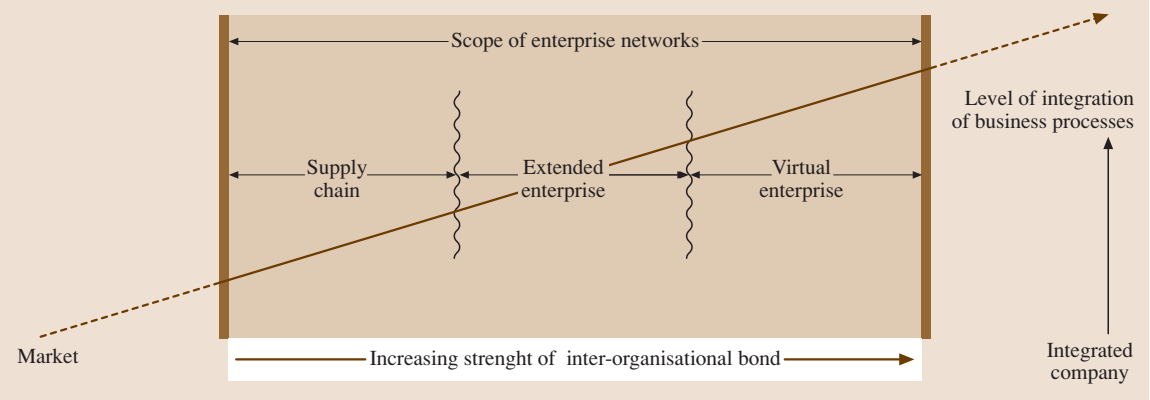


Fig. 15.72 Types of collaborations within enterprise networks

relationship into clearly defined and distinct compartments of SC, EE, and VE. Hence, the types of bilateral relationships depicted in Fig. 15.72 indicate (and should be taken as representing) a *trend* in the closeness of relationships between the cooperating enterprises. By working together, the level of confidence may improve and the relationship may evolve to a closer cooperative level. It is not possible to define a single criterion or demarcation when the collaboration evolves from an SC type of relationship to EE, or from EE to VE. To a large extent the boundary could vary with the product sector as well as the market environment. Indeed, there can be a wide region between two classifications where arguments will persist regarding the name of that collaboration. It is also worth pointing out that Fig. 15.72 does not strictly represent evolution in collaboration, as the following sections will make clear.

While the concept of supply chains has been well established, it is the emergence of information technology (IT), which deals with the sharing or exchange of information between two parties, and communication technologies (CT), which focus on the tools required for the actual transfer of information between any two parties), that has expedited the nature and scope of collaboration to new higher levels. In this section the terms *supply chains* and *distribution chains* are used interchangeably. Even though the mechanisms of setting up supply chains (materials/products sourcing) and distributions chains (customers sourcing and distribution center location) could be different, the only difference perceived by us is the perspective from which one is looking at. Delivery from enterprises upstream to one's operations is via supply chains (a *pull system*) and the distribution of completed products will require distribution chains (a *push system*). The logistics involved in either case is similar. The formation of extended enterprise and virtual enterprise are very recent developments. Therefore, *extended and virtual enterprises are merely new paradigms reflecting the extent to which the information systems of the collaborating enterprises are integrated with one another and the way in which they actually communicate and collaborate with one another. In other words, extended and virtual enterprises are different (and more sophisticated) manifestations of the supply chains. Hence, most of the underlying principles and operational issues prevalent in a supply chain will be present in extended and virtual enterprises. Indeed, the supply chain between the collaborating enterprises has to be setup before they* switch *to extended or virtual enterprise mode* [15.176].

Theoretically extended and virtual enterprises could occur in any manufacturing environment. However, current trends indicate that certain environments are more conducive to their formation, where the partners can clearly identify the resulting benefits. Within traditional manufacturing environments, one could locate the sectors, based on the product complexity and lot sizes, in which the formation of extended and virtual enterprises will be more prevalent. Figure 15.73 attempts to approximately locate such regions. In this diagram, when we talk about the complexity of a product, we mean not only the number of components it has or how complex it is to assemble but, more importantly, how technologically sophisticated its subcomponents are and to what extent they will require external core competencies in their own right for their manufacture. Typical examples of technologically complex components could be graphics cards for personal computers (PCs), fuel injection systems for automobiles, etc.

ICT (IT and CT) technologies can be seen as major enablers for modern enterprise collaborations. Without appropriate ICT-based approaches and infrastructures,

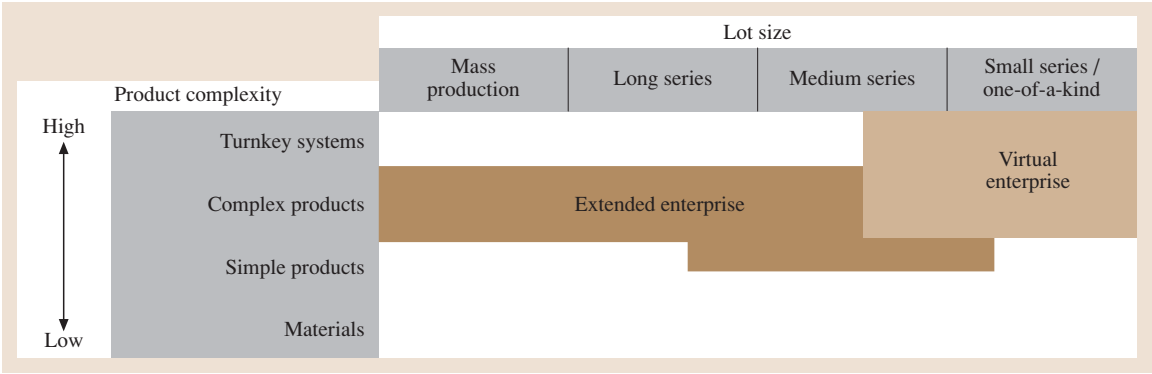


Fig. 15.73 Locating EE and VE within traditional manufacturing environments

cooperative relationships between enterprises, as dictated by modern-day market conditions, would not be possible. By improving the transparency of interorganizational infrastructures and implementing the advanced organizational approaches, cooperative and flexible solutions can become feasible. These technologies have now become a prerequisite for highly economic enterprise cooperations. Whereas classical IT solutions supported mainly the rationalization of the production process, ICT solutions of the 1990s tried to reduce the transaction costs as well, thus forming an integral part of the collaboration. Without appropriate ICT-based approaches and infrastructures, effective cooperative relationships between enterprises would not be possible. This has given the manufacturers more choices in terms of in-house product development and outsourcing. By improving the transparency of interorganizational product development as well as order processing ICT enables cooperative and flexible solutions.

Various studies have shown that ICT could reduce the coordination costs that result from outsourcing activities and therefore stimulate cooperation between firms [15.177]. Important IT developments that enable interorganizational cooperation also include the development of standards for exchanging information, e.g., EDIFACT (electronic data interchange for administration, commerce and transport) and STEP (standard for the exchange of product data) [15.178], the reduction of prices for computational power of processors, and the interconnectivity of communication networks, resulting in the Internet and the like. *Williamson*'s transaction cost theory also predicts that weighing economies of scale against transaction costs leads to the outsourcing of activities that have been carried out so far inside the organization [15.175].

Enterprise collaborations have also driven the requirements and developments in the information and telecommunication technologies; for example, developments in EDI (electronic data interchange) technology and STEP protocols to facilitate the transfer of design data and indeed product models between partners has redefined the concept of collaboration to new levels.

Supply Chain Type of Collaboration. A supply chain is a set of activities by which several enterprises (termed nodes) agree to contribute their expertise towards the completion and supply of a common end-product. A simplified model of a supply chain is depicted in Fig. 15.74. Each node in the supply chain acts as a customer as well as a supplier. In the customer mode, it receives (buys) unfinished goods from upstream suppliers, uses its core competence to add value to the product, and passes (sells) them onto the next node downstream in the chain. In traditional manufacturing, the supply chain could span raw-materials suppliers right through to the consumer of the end-product, thus encompassing all the intermediate activities of manufacturing, storage, distribution, and delivery. The *supply-side* limit of the supply chain is dictated by the criticality of the raw materials. The term *raw materials* is relative and highly dependent on the market and product sector. It also includes the issues of how readily the raw materials are available in the marketplace in which the supply chain is operating. The distribution side of the supply chain need not be restricted to the final consumer. Indeed, with the emerging environmental and associated recycling issues, enterprises involved in the disposal of the product at the end of its useful life may well form part of this supply chain. A supply chain need not be a sequential set of nodes. In fact, in real life it will often take the form of an enterprise network.

The concept of supply chains evolved long before the advent of ICT. Basic research related to supply-chain management (SCM) has been carried out since the 1960s. However, it was in the 1980s that market pressures on producers and suppliers, and the emerging ideas of just-in-time (JIT) manufacturing, gave impetus to the study of SCM and its wider acceptance; thus it formally became an integral part of manufacturing and operations research jargon. SCM entails coordination with customers and suppliers. In order to operate efficiently all nodes across the supply chain must operate in a synchronous mode, providing a rapid and high-quality response to the events.

In the dynamics of real life, perturbations to planned events will *always* occur. Unplanned events (changes in the customer order, delayed delivery of materials, machine breakdown, etc.) *will* cause deviations from scheduled activities. Furthermore, it is the accumulation of perturbations across the supply chain that can make the whole venture very inefficient. As an example, consider two cases:

1. The customer changes the order composition: this effect will ripple and amplify upstream in the supply chain as other nodes sequentially readjust their operations to incorporate this change.
2. Materials are not delivered according to agreed schedules: this effect will ripple and amplify downstream in the supply chain as other nodes sequentially readjust their operations to incorporate this change. Like the market, the production floor is

also dynamic. Therefore, the same effect will be observed if there was production perturbation, say due to a machine breakdown at one of the nodes.

In either case, the only way to counteract either of these effects is to have larger inventories. Perturbations occurring at two different nodes (due to different reasons) will further amplify the *noise* in the system. Hence, information lags, delivery lags, demand volatility, unsynchronized ordering, over- and underordering, lumpy ordering, and chain accumulations will all contrive to make demand become more volatile further along the supply chain.

The primary requirements of the operational phase of any supply chain are the minimization of inventory and lead times across the whole chain. To achieve this:

- Partners in the supply chain need to have a clear understanding (contractual or otherwise) as to what is expected from each partner. This is also true for the respective expectations.
- Both material and information flow systems need to be streamlined.
- Various functions within a node, such as marketing, sales, purchasing, production planning, and production control, must communicate effectively with one another.
- Information exchange among the nodes must occur efficiently for a supply chain to operate effectively.
- Information and decision support systems at nodes must be able to respond dynamically to meet the ever-changing needs and communicate accordingly to the affected nodes of the supply chain.
- The industry is moving from being product led to being customer led, and those in the supply chain will need to change their ways of operating to respond to this change.
- Issues such as quality management and continuous improvement are often part of the contractual agreement when the supply chain is set up.

This can only be achieved through excellence in communications among the supply-chain partners. This requires an open, scalable, flexible communications solution that provides both seamless communications and the ability to adapt rapidly to changes in supply-chain structure. The Internet is one possible solution because it has the distinct advantage of being a completely open, global communications system, with deep market penetration.

As we will see in the following sections, the emergence of affordable Internet-based ICT technologies is pushing the frontiers of enterprise collaboration to new levels and, as a result, the emergence of new paradigms from the supply chain type of collaboration: extended and virtual enterprises.

Extended Enterprise Type of Collaboration. The traditional view of business organizations is no longer valid [15.179–181]. Instead of speaking of industrial collaboration, however, one uses the term *extended enterprise* when referring to a new paradigm for manufacturing. The extended enterprise is a term frequently used in today's business literature to reflect the high level of cooperation between organizations. *Browne* et al. [15.179] argue that computer-integrated manufacturing (CIM) will enable interenterprise networking across the value chain. Manufacturing systems are no longer confined to a single factory, but cross enterprise boundaries. Integration of operations of independent organizations with the operations of suppliers and customer can result in extended enterprises. Furthermore, the market sector is not restricted to manufacturing but many other business areas such as financial services, distribution, and information services have formed closer relationships that can be termed as extended enterprises.

Extended enterprises are evolutionary in nature. Let us take an example. Two organizations have known each other and conducted business in a supply chain for some time. During this period of collaboration a sufficient level of trust has developed to automate the sharing of day-to-day operational data. This integration will be preceded by the realization of the importance of each organization of the other in its business plans. Therefore, each of the organizations will be prepared, if necessary, to invest in modern ICT tools for the effortless sharing of information. The fact that they are willing to invest in their collaboration will imply that they are committed to a long-term relationship. It is this seamless exchange of relevant operational information on top of an existing long-term (and successful) relationship that distinguishes the extended enterprise form other forms of long-term collaboration such as a supply chain. It should also be noted that ICT are enabler technologies and a necessary (though not sufficient) condition for an extended enterprise to exist. It is the integration of the respective information and decision systems and the respective production processes that link them closely enough (within the agreed bounds) to be analogous to the be-

havior of a single enterprise operating within the *four walls*.

The extended enterprise can be regarded as a kind of *enterprise* which is represented by all those organizations or parts of organizations, customers, suppliers, and subcontractors, engaged collaboratively in the design, development, production, and delivery of a product to the end user [15.182, 183]. Key suppliers become almost a part of the principal company and its information infrastructures, with frequent exchange of status information. This is echoed succinctly in the definition of extended enterprise given in *Jagdev* et al. [15.184] as "the formation of closer coordination in the design, development, costing and the coordination of the respective manufacturing schedules of cooperating independent manufacturing enterprises and related suppliers." The key term in this definition is the *coordination of the respective manufacturing schedules*. This coordination of respective schedules, which includes not only the production schedules but also dispatch, transportation/delivery, and receipt notifications, which is supposed to be performed seamlessly through the use of ICT technologies, is a necessary condition for the formation of an extended enterprise, because only then can one truly realize the integration of respective IT infrastructures, which again is a necessary condition for the formation of an extended enterprise.

The baseline for an extended enterprise is always two or more willing enterprises which have chosen to concentrate on their core competencies and wish to extend their activities into other enterprises to increase their competitiveness by achieving cost-, time- or quality-related advantages regarding their respective offerings. Extending activities implies that an enterprise is enhancing its existing capabilities or adding additional facilities, by outsourcing, which have not been at its disposal so far. Outsourcing encourages both the manufacturer and its suppliers' competitive ability and enhances their mutual dependency.

Taken from this perspective, we can identify the following as major characteristics of the extended enterprise:

- The partners in the extended enterprises are willing to form long-term relationships and treat each other as business partners. Each partner understands and accepts other's requirements and priorities.
- Within the scope of collaboration, partners share vision and work towards shared goal.
- Decisions are jointly arrived at by making best use of the competencies among the partners.
- The primary mode of communication and sharing of information between the collaborating enterprises will always be through telecomputing. It is therefore important to have available advanced ICT tools to support the extended enterprise [15.185].
- The efficiency of the extended enterprise is greatly determined by the speed and efficiency with which information can be exchanged and managed among business partners. Efficient collaborative engineering, production, and logistics require effective electronic management of engineering and production information. Thus it is important that the participating enterprises have sufficiently sophisticated IT and decision support tools and mechanisms to make this integration possible. It is also important to have the maximum degree of compatibility among partners' IT systems.
- Day-to-day communications between the respective IT systems of two enterprises will always be real time and online and without human intervention; for example, if there are production schedule perturbations in the *customer* enterprise, these changes will (and should) be automatically communicated to the *supplier* enterprise IT systems, thus triggering the processes necessary for updating of production schedules at the *supplier* end.
- Extended enterprise can occur between any two enterprises across the value chain of any product or service. Enterprises across the whole value chain can be involved in the extended enterprise. However, the concept of the whole value chain is neither a necessary nor a relevant condition for the formation of extended enterprises. If this were the case, we would enter a very slippery slope of defining the

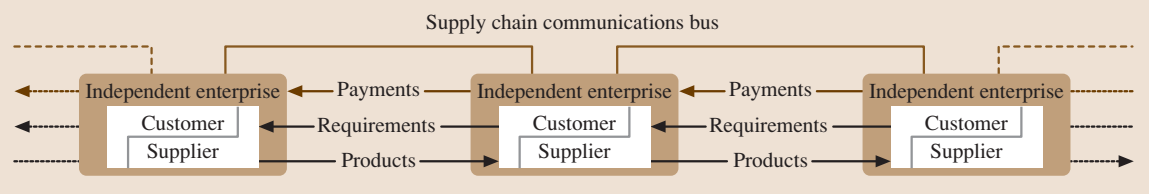


Fig. 15.74 Simplified model of a supply chain

boundaries of the value chain; for example, why restrict the starting point of the value chain to raw materials; why not extend it to their mining? Furthermore, *raw materials* for one value chain could very well be a *finished product* for another value chain, e.g., ball bearings, electric motors, transformers/power supply units, etc.
- Technology permitting, an extended enterprise can take the form of a complex enterprise network in which each enterprise can be seen as a node.
- The relationship between a set of nodes in an extended enterprise can be hierarchical or nonhierarchical.

Virtual Enterprise Type of Collaboration. Unlike the extended enterprise formation in which, more often than not, existing supplier/buyer (or supply chain) relationships are strengthened through the usage of ICT technologies, the availability of ICT technologies can initiate the formation of entirely new enterprise networks.

A virtual enterprise is one manifestation of organizational response to the dynamism and globalization of today's markets. The baseline for a virtual enterprise is the customer needs. These needs can be extensive and unique (e.g., a large project-based contract) or small but with numerous variations; for example a number of complementary companies specializing in the repair and maintenance of household items may form a virtual enterprise to provide a comprehensive service to their potential customers. These services might include the maintenance of house structure, all forms of energy supplies, telecommunication and entertainment links, repair/maintenance of household durables such as cookers, washing machines, refrigeration equipment, and the recycling and waste disposal. Each of these services will require unique core competencies. Thus, several small specialist companies can increase their potential customer base by pooling their competencies. The attractiveness for the customer of such an enterprise will be that there will be only a single contact point for most of their household-related problems.

From the above argument we can note that there needs to be pooling of more than one core competence for the formation of virtual enterprise. Taken literally, there is nothing new in small companies joining forces to strengthen their marketability. However, it is the availability of ICT technologies that has given small enterprises entirely new platforms to collaborate efficiently to supply effective product/service. Thus, it should be noted that ICT are enabling technologies and a necessary (though not sufficient) condition for a virtual enterprise to exist.

Business partners in virtual enterprises can retain the individual agility of the consortium members to undertake their own business operations and quiet possibly participate in other EE or VE type of projects simultaneously [15.184, 186]. As information and communications technologies overcome the constraints of time and distance, it has become possible to create virtual organizations. Those independent companies, temporarily linked by ICT networks, share skills and costs, and can access each other's markets.

Some authors define virtual enterprises as temporary networks of independent companies engaged in providing a product or service. *Childe* [15.160] defined the evolution of virtual enterprise as: and "... in the extreme, the principal company may have no premises and may not perform what is generally seen as manufacturing work – it may only consist of functions such as design and production management, acting as the co-ordinator of activities in the cooperating companies. In this case, the company becomes a virtual enterprise, appearing to its customers as a supplier of goods and services, but with no internal production activities." In such a scenario, virtual enterprises are often project based and usually operate in niche markets. Thus they form, reform, and dissolve based on the market dynamics and the opportunities it provides.

Forbairt [15.187] depicts the virtual enterprise as a response to the speed and globalization of the digital age. It is an enterprise that may have no physical head office or center and very few full time workers, and that exists as a combination of specific skills and competencies of individuals or enterprises. A virtual enterprise may be set up with the objective of providing one particular type of product or service. When the market for that product or service declines, the virtual enterprise dissolves, its members finding new partners to pursue new opportunities. In the global village, companies who survive and prosper are agile, flexible, and adaptable and constantly seek to maximize their unique advantages.

Compared to an original enterprise, according to *Scholz* [15.188], a virtual enterprise is always characterized by the absence of specific physical attributes/features such as a common administration or a common legal status. These features are replaced by the application of sophisticated information and communication infrastructures and mutual confidence (common understanding).

Skyrme [15.189] emphasizes the growth of networking, both human and technological, and argues that this

is creating a virtual world with virtual products and services, virtual workplaces, and virtual organizations. The virtual products and services are produced, delivered and sold through electronic networks, e.g., telebanking and telemarketing, significantly reducing the costs of many activities and creating opportunities to reach distant markets easily. The virtual workplaces include teleworkers, who work in a location-independent manner (sometimes at home), and flexible offices. Virtual organizations work in teams and cooperate across company boundaries to create the organization necessary for and appropriate to specific projects, thereby gaining significant flexibility in the use of people. This virtual world is the broad social environment for the establishment of the virtual enterprise in manufacturing.

The intelligent manufacturing systems (IMS) project [15.190] defines the virtual enterprise as the next generation of manufacturing enterprise, which consists of a globally distributed assembly of autonomous work units linked primarily by the goal of profitably serving specific customers and operating in an environment of abrupt and often unanticipated change.

Davidow and *Malone* [15.180] echo a similar sentiment when they suggest that networks of cooperating manufacturers are the industrial strategy for structuring and revitalizing corporation for the 21st century. They term these networks *virtual corporations*. We believe that this definition of virtual corporation is rather generic, as it also encompasses many of the principles of electronic commerce. The virtual enterprise described in this section is a more formal definition depicting the nature of relationship between two (or more) organizations engaged in providing a real product or service, as opposed to some virtual product. Our interpretation of virtual enterprise can be considered as a subset of virtual corporation as defined by *Davidow* et al. In their book *Virtual Corporation* [15.180] they present their vision of 21st century industries, which will be built around "a new kind of product, delivering instant customer gratification in a cost-effective way." These products have a very rich service component that is often more important than the tangible characteristics of the product. They can be produced in several locations and offered in a great number of models and formats. Davidow et al. call these products *virtual products*. They believe that a manufacturing company will not be an isolated facility of production, but rather a node in the complex network of suppliers, customers, engineering, and other service functions. The real-time adaptation of the virtual product to customer needs requires the virtual corporation to maintain integrated and ever-changing data files on customers, products, and production and design methodologies. They therefore speak of it in terms of patterns of information and relationships that will appear less as a discrete enterprise and more as an ever-varying cluster of common activities of suppliers and their downstream customers in the midst of a vast fabric of relationships. These relationships will be built on principles such as *joint destiny, trust, and sharing information*.

The virtual corporation of Davidow and Malone is described as almost edgeless, with permeable and continuously changing interfaces between company, supplier, and customers. Nevertheless, Davidow and Malone stress the importance of brand names and product identity. A virtual corporation is identified by the activities carried out and the products delivered. In fact, a virtual corporation is defined through the product or product line it produces.

Møller [15.191] defines the virtual enterprise from the supply-chain point view. This concept is used to characterize the global supply chain of a single product in an environment of dynamic networks of companies engaged in many different relationships. The companies in a virtual enterprise coordinate their internal systems with other systems in the supply chain, and simultaneously participate in other virtual enterprises and adapt to changes.

In principle, small and medium-sized companies participating in a virtual enterprise gain access to the resources of a large organization while retaining the agility and independence of a small one. *Skyrme* [15.189] suggests that the following benefits may be obtained through the construction of a virtual enterprise:

- Access to a wide range of specialized resources
- Presentation a unified face to large corporate buyers
- That individual members retain their independence and continue to develop their core competencies
- Reshaping of the enterprise and changing of members according to the project or task in hand
- That there is no need to worry about *divorce settlements* as in formal joint ventures

Therefore, a *virtual enterprise can be defined as a network of independent organizations that jointly form an entity committed to provide a product or service*, because from the customer's perspective as far as that product/service is concerned, these independent organizations, for all practical and operational purposes, are *virtually* acting as a single entity/enterprise. Taken

from this perspective, we can identify the following as major characteristics of the virtual enterprise:

- The (two or more) partners in the virtual enterprises are individuals and independent companies who come together and form a *temporary* consortium to exploit a particular market opportunity.
- Within the scope of collaboration, partners share vision and work towards shared goal.
- Partners in virtual enterprises make extensive use of ICT technologies for communication and sharing information. Most of the day-to-day information exchange among the partners will almost always be automatic and without human interference.
- Virtual enterprises assemble themselves based on cost effectiveness and product uniqueness without regard to organization size or geographic location.
- Unlike SCs or EEs, virtual enterprises, once formed, will have a unique dynamics, new identity, and quiet possibly a new name.
- The efficiency of the virtual enterprise is greatly determined by the speed and efficiency with which information can be exchanged and managed among business partners. Efficient collaborative engineering, production, and logistics require effective electronic management of engineering and production information. Thus it is a prerequisite that participating enterprises have sufficiently sophisticated IT and decision support tools and mechanisms to make this integration possible.
- Virtual enterprises pool costs, skills, and core competencies to provide world-class solutions that could not be provided by any one of them individually. Therefore, virtual enterprises often focus on complete products or solutions as opposed to providing partial solutions in a value chain.
- Decisions are jointly arrived at by making best use of the competencies among the partners.
- Virtual enterprises will often be complex networks in which each enterprise can be seen as a node.
- The relationship between a set of nodes in a virtual enterprise will mostly be nonhierarchical in nature.

References

15.1 H.S. Jagdev, K.-D. Thoben: Anatomy of enterprise collaborations, J. Prod. Planning Control **12**(5), 437–451 (2000)

15.2 B. Wu: *Manufacturing System Design and Analysis*, 2nd edn. (Chapman Hall, New York 1994)

15.3 DIN: *DIN 8580: Manufacturing Processes – Terms and Definitions* (Beuth, Düsseldorf 2006)

15.4 H.C. Kazanas, G.E. Baker, T. Gregor: *Basic Manufacturing Processes* (McGraw-Hill, New York 1981)

15.5 R.L. Francis, J.A. White: *Facility Layout and Location: An Analytical Approach* (Prentice-Hall, Englewood Cliffs 1974)

15.6 B. Kirwan, L. Ainsworth (Eds.): *A Guide to Task Analysis* (Taylor and Francis, Englewood Cliffs 1992)

15.7 M. Imai: *Kaizen – The Key to Japan's Competitive Success* (McGraw-Hill, New York 1986)

15.8 J.H. Heizer, B. Render: *Production and Operations Management*, 3rd edn. (Allyn and Bacon, Boston 1994)

15.9 E. Frese: *Grundlagen der Organisation* (Gabler, Wiesbaden 1981)

15.10 R. Likert: *New Patterns of Management* (McGraw-Hill, New York 1961)

15.11 N. Slack, S. Chambers, C. Harland, A. Harrison, R. Johnston: *Operations Management*, 2nd edn. (Pearson, London 1998)

15.12 H. Wildemann: *Die modulare Fabrik: Kundennahe Produktion durch Fertigungssegmentierung* (Gesellschaft für Management und Technologie, München 1994), in German

15.13 H. Kühnle: L'entreprise fractale. In: *La Modélisation Systémique en Entreprise*, ed. by C. Braesch, A. Haurat (Pôle productique Rhône-Alpes, Paris 1995)

15.14 G. Schuh, K. Millarg, A. Göransson: *Virtual Factory* (Hanser, Munich 1998)

15.15 J.J. Modder, C.R. Phillips, E.W. Davis: *Project Management with CPM/PERT and Precedence Diagramming* (Van Nostrand, New York 1983)

15.16 PMBOK Guide: Project Management Institute (PMI): A Guide to the Project Management Body of Knowledge, Pennsylvania (2000)

15.17 J.B. Barney: Firm resources and sustained competitive advantage, J. Manage. **17**(1), 99–120 (1991)

15.18 I. Dierickx, K. Cool: Asset stock accumulation and sustainability of competitive advantage, Manage. Sci. **35**(12), 1504–1513 (1989)

15.19 R.P. Rumelt: Towards a strategic theory of the firm. In: *Competitive Strategic Management*, ed. by R.B. Lamb (Prentice Hall, Englewood Cliffs 1984)

15.20 J.A. Schumpter: *The Theory of Economic Development* (Harvard Univ. Press, Cambridge 1934)

15.21 E. Penrose: *The Theory of the Growth of the Firm* (Wiley, New York 1959)

15.22 B. Wernerfelt: A resource-based view of the firm, Strategic Manage. J. **5**, 171–180 (1984)
15.23 R. Daft: *Organization Theory and Design* (West, New York 1983)
15.24 J.B. Barney, M. Wright, D.J. Ketchen Jr.: The resource-based view of the firm: Ten years after 1991, J. Manage. **27**, 625–641 (2001)
15.25 R.M. Grant: The resource-based theory of competitive advantage: Implications for strategy formulation, Calif. Manage. Rev. **33**(3), 114–135 (1991)
15.26 M.E. Porter: *Competitive Strategy* (Free, New York 1980)
15.27 M.E. Porter: Towards a dynamic theory of strategy, Strategic Manage. J. **12**, 95–117 (1991), (special issue)
15.28 C.K. Prahalad, G. Hamel: The core competence of the corporation, Harvard Bus. Rev. **68**(3), 79–91 (1990)
15.29 J.S. Harrison, M.A. Hitt, R.E. Hoskisson, R.D. Ireland: Resource complementarity in business combinations: Extending the logic to organizational alliances, J. Manage. **27**, 679–690 (2001)
15.30 B. Kogut, U. Zander: What firms do?: Coordination, identity, and learning, Organ. Sci. **7**, 502–518 (1996)
15.31 R. Amit, P.J.H. Schoemaker: Strategic assets and organizational rent, Strategic Manage. J. **14**(1), 33–46 (1993)
15.32 D.J. Teece, G. Pisano, A. Shuen: Dynamic capabilities and strategic management, Strategic Manage. J. **18**(7), 509–533 (1997)
15.33 J.H. Dyer, H. Singh: The relational view: Cooperative strategy and sources of interorganizational competitive advantage, Acad. Manage. Rev. **23**(4), 660–679 (1998)
15.34 Y.J. Shi, M.J. Gregory: International manufacturing networks – to develop global competitive capabilities, J. Oper. Manage. **16**, 195–214 (1998)
15.35 R.R. Nelson, S.G. Winter: Evolutionary theorizing in economics, J. Econ. Perspect. **16**(2), 23–46 (2002)
15.36 R. Makadok: Toward a synthesis of the resource-based and dynamic capability views of rent creation, Strategic Manage. J. **22**, 387–401 (2001)
15.37 M. Fiol: Revisiting an identity-based view of sustainable competitive advantage, J. Manage. **27**(6), 691–699 (2001)
15.38 K. Eisenhardt, J. Martin: Dynamic capabilities: What are they?, Strategic Manage. J. **21**, 1105–1121 (2000)
15.39 M.A. Lewis: Analysing organizational competence: Implications for the management of operations, Int. J. Oper. Prod. Manage. **23**(7), 731–756 (2003)
15.40 C.E. Helfat, M.A. Peteraf: The dynamic resource-based view, Capability Lifecycles **24**, 997–1010 (2003)
15.41 R.H. Hayes, S.C. Wheelwright: *Restoring Our Competitive Edge – Competing Through Manufacturing* (Wiley, New York 1984)
15.42 J.S. Srai, M.J. Gregory: Supply Chain Capability Assessment of Global Operations, Euroma Conference Proc Budapest (2005)
15.43 J.B. Barney: How a firm's capabilities affect boundary decisions, Sloan Manage. Rev. **40**(3), 137–148 (1999)
15.44 F.T.S. Chan, H.J. Qi: Feasibility of performance measurement for supply chain: a process based approach and measures, Integr. Manuf. Syst. **14**(3), 179–190 (2003)
15.45 J.S. Srai, D. Fleet, Y. Shi, M.J. Gregory: Identification of Supply Chain Capabilities in International Supply Networks, Euroma Conference Proc INSEAD (2004)
15.46 G. Spur (Ed.): *Handbook of Production Engineering, Volume 6: Factory Operations* (Hanser, Munich 1993), in German
15.47 G. Ropohl: *General Technology – A System Theory of Technique*, 2nd edn. (Hanser, Munich 1999), in German
15.48 H.-G. Riehle, P. Rinza, H. Schmitz: *Systems Engineering in Business and Administration, Volume 1: Fundamentals and Methods* (VDI, Düsseldorf 1978), in German
15.49 J. Baetge: *Operational System Theory* (Westdeutscher Verlag, Opladen 1974), in German
15.50 K. Mertins, W. Süssenguth, R. Jochem: An object oriented method for integrated enterprise modelling as a basis for enterprise coordination. In: *Enterprise Integration Modelling, Proceedings*, ed. by C.J. Petrie (MIT Press, Cambridge 1992) pp. 249–258
15.51 C. v. Uthmann, J. Becker: Guidelines of modelling (GoM) for business process simulation. In: *Process Modelling*, ed. by B. Scholz-Reiter, H.-D. Stahlmann, E. Nethe (Springer, Berlin 1999) pp. 100–116
15.52 VDI: *Assembly and Handling: Handling Functions, Handling Units; Terminology, Definitions and Symbols, VDI 2860* (VDI, Düsseldorf 1990), in German
15.53 A.-W. Scheer: *Business Process Automation: ARIS in Practice, with 6 Tables* (Springer, Berlin 2004)
15.54 F. Vernadat: *Enterprise Modelling and Integration – Principles and Applications* (Chapman Hall, London 1996)
15.55 IFIP-IFAC Task Force: GERAM – Generalised Enterprise Reference Architecture and Methodology. Version 1.6.3, March 1999, Annex to ISO WD15704, Requirements for Enterprise-Reference Architectures and Methodologies (Düsseldorf 1999)
15.56 ISO WD15704: Industrial Automation Systems - Requirements for Enterprise Reference Architectures and Methodologies (Düsseldorf 1999)
15.57 CIMOSA Association: CIMOSA: Open System Architecture for CIM: Technical Baseline, Version 3.2 (1994)
15.58 M. Zelm (ed.): *CIMOSA: A Primer on Key Concepts, Purpose and Business Value* (CIMOSA Association e.V., Stuttgart 1995)

15.59 ESPRIT Consortium AMICE (ed.): *CIMOSA: Open System Architecture for CIM. Research Reports ESPRIT, 2nd, revised and extended edition* (Springer, Berlin 1993)

15.60 H. Kühnle, J. Braun, M. Hüser: Produzieren in turbulentem Umfeld. In: *Aufbruch zum Fraktalen Unternehmen*, ed. by H.-J. Warnecke (Springer, Berlin Heidelberg 1995)

15.61 H. Kühnle, G. Spengler: Approaches to the fractal company, IO Manag. Z. **62**(4), 66–71 (1993), in German

15.62 W. Dangelmaier, R. Bachers: Parametrized Simulation System SIMULAP, AMSE Modelling and Simulation (Minneapolis 1984)

15.63 P.S. Chen: The entity-relationship model – toward a unified view of data, Trans. Database Syst. **1**(1), 9–36 (1976)

15.64 J. Rumbaugh, I.G. Jacobson, G. Booch: *The Unified Modelling Language Reference Manual: UML, [covers UML 2.0]*, 2nd edn. (Addison-Wesley, Boston 2005)

15.65 Rational Software (eds.): *Rational Unified Process: Best Practices for Software Development Teams* (Rational Software, Lexington 1998)

15.66 OMG SysML home page: www.omgsysml.org (last visited July 2007)

15.67 P. Kruchten: *The Rational Unified Process – An Introduction* (Addison-Wesley, Reading 2003)

15.68 T.E. Vollmann, W.L. Berry, D.C. Whybark: *Manufacturing Planning and Control Systems* (Irwin, Chicago 1992)

15.69 H. Luczak, W. Eversheim: *Produktionsplanung und Steuerung* (Springer, Berlin 2006), in German

15.70 J. Orlicky: *Material Requirements Planning* (McGraw-Hill, New York 1975)

15.71 H. Kühnle: Integration of CAPP and PPC – The PPC Part. In: CIRP International Institute for Production Engineering Research: CIRP Annals – Vol. 41/1/1992, S. 15–18

15.72 B. Ritter: *Enterprise Resource Planning (ERP)* (Mitp-Verlag, Heidelberg 2005)

15.73 J. Kenworthy: *Planning and Control of Manufacturing Operations* (Woodhead/IOM, Cambridge 1998)

15.74 G. Zäpfel: *Produktionswirtschaft* (de Gruyter, New York 1992)

15.75 B. van Hezewijk, M. van Assen: Coordination in Chains and Networks, EUROMA Conference Proc Budapest (2005) pp. 875–886

15.76 N.N.: Product Specification, infor business solutions AG (2003), in German

15.77 O. Wight: *MRP II: Unlocking America's Productivity Potential* (CBI, Boston 1981)

15.78 H.A. ElMaraghy: Flexible and reconfigurable manufacturing systems paradigms, Int. J. Flex. Manuf. Sys. **17**, 4 (2005)

15.79 M.L. Pinedo: *Planning and Scheduling in Manufacturing and Services* (Springer, New York 2005)

15.80 K.R. Baker: *Introduction to Sequencing and Scheduling* (Wiley, New York 1984)

15.81 http://www.pabadis-promise.org/ (last accessed 2007-05-22)

15.82 J. Ashayeri, R. Kampstra: Realities of Supply Chain Collaborations, EUROMA Conference Proc Budapest (2005) pp. 339–348

15.83 M. Baratt, A. Olivera: Exploring the experiences of collaborative planning initiatives, Int. J. Phys. Distrib. Logist. Manage. **31**(4), 266–289 (2001)

15.84 G. Fliedner: *CPFR: An emerging supply chain tool*, Ind. Manage. Ind. Sys. **103**(1), 14–21 (2003)

15.85 M. Howard, J. Miemcyk, G. Stone, A. Graves: Coping with Collaborative ICT Ventures: Implementing the responsive Automotive Supply Network, EUROMA Conference Proc Glasgow (2006) pp. 985–994

15.86 W.J. Hopp, M.L. Spearman: *Factory Physics: Foundations of Manufacturing Management*, 2nd edn. (McGraw-Hill, New York 2001)

15.87 J. Jiao: Generic bill-of-materials-and-operations for high-variety production management, Concurrent Eng-Res. A **8**(4), 297–322 (2000)

15.88 Manufacturing Enterprise Solutions Association: Homepage, www.mesa.org, (last visit February 2006)

15.89 W. Shen, D. Norrie: Agent-based systems for intelligent manufacturing – A state-of-the-art survey, Int. J. Knowl. Inform. Sys. **1**(2), 129–156 (1999)

15.90 A. Lüder, J. Peschke, D. Reinelt: Possibilities and Limitations of the Application of Agent Systems in Control, 12th Int. Conf. Concurrent Enterprising (ICE2006) (Milan 2006) pp. 149–156

15.91 http://www.orin.jp/ (last accessed June 2005)

15.92 http://www.opcfoundation.org/ (last accessed June 2005)

15.93 D. Yui, S. Sakakibara: Database Application Software for Production Management Using RAO, Proceedings (III) of SICE (Society of Instrument and Control Engineering) Integration Division Annual Conference (2002) pp. 297–298

15.94 IBM Japan, Ltd.: OpenMES Specification Ver.1.0 Draft Alpha (May 25. 1999)

15.95 T. Kimura, Y. Kanda: Development of a remote monitoring system for a manufacturing support system for small and medium-sized enterprises, Comput. Ind. **56**, 3–12 (2005)

15.96 T. Tsukikawa, T. Hironaka, T. Otsuka, M. Mizukawa, Y. Ando: Research on Remote Control of AIBO Using ORiN, Proceedings of SICE (Society of Instrument and Control Engineering) Integration Division Annual Conference (2003) pp. 21–22

15.97 E. Zahn, R. Dillerup, S. Foschiani: Ansätze zu einem ganzheitlichen Produktionsmanagement. In: *Ganzheitliche Unternehmensführung*, ed. by H.D. Seghezzi (Schäffer-Poeschel, Stuttgart 1997), in German

15.98 R.D. Stacey: *Complexity and Creativity in Organizations* (Berrett-Koehler, San Francisco 1996)

15.99 T.S. Kuhn: *The Structure of Scientific Revolutions* (Univ. of Chicago Press, Chicago 1962)
15.100 J. Womack, D. Jones, D. Roos: *The Machine that changed the World* (Rawson Associates, New York 1991)
15.101 E.M. Goldratt: *Essays on the Theory of Constraints* (North River, Great Barrington 1988)
15.102 J. Liker: *The Toyota Way: 14 Management Principles from the World's Greatest Manufacturer*, 1st edn. (McGraw-Hill, New York 2003)
15.103 T. Ohno: *Toyota Production System: Beyond Large-Scale Production* (Productivity Press, Cambridge 1988)
15.104 D.J. Bennett: *Lean Production and Work Organization* (MCB Univ. Press, Bradford 1996)
15.105 P.T. Kidd: *Agile Manufacturing: Forgoing New Frontiers* (Addison-Wesley, Reading 1994)
15.106 S.L. Goldman, R.N. Nagel, K. Preiss: *Agile Competitors and Virtual Organizations* (Van Nostrand Reinhold, New York 1995)
15.107 K. Pawar, S. Sharifi: Product development strategies for agility. In: *Agile Manufacturing: The 21st Century Strategy*, ed. by G. Gunasekaran (Elsevier, New York 2001)
15.108 N. Okino: Bionic Manufacturing Systems, Conference on Flexible Manufacturing Systems, Past, Present-Future, ed. by J. Peklenik (Faculty of Mechanical Engineering, Ljubljana 1993)
15.109 K. Ueda: Biological Manufacturing Systems, 1996 NGMS International Conference (Irvine 1996)
15.110 A. Lüder, L. Ferrarini, C. Veber, J. Peschke, A. Kalogeras, J. Gialelis, J. Rode, D. Wünsch, V. Chapurlat: Control Architecture for Reconfigurable Manufacturing Systems: The PABADIS'PROMISE approach, 11th IEEE Int. Conf. on Emerging Technologies and Factory Automation (ETFA'06) (Prague 2006) pp. 545–552
15.111 A. Koestler: *The Ghost in the Machine* (Hutchinson, London 1971)
15.112 H. van Brussels: Holonic Manufacturing Systems – The Vision Matching the Problem, First European Conference on Holonic Manufacturing Systems (Hanover 1994)
15.113 J. Christensen: Holonic Manufacturing Systems: Initial Architecture and Standards Directions, First European Conference on Holonic Manufacturing Systems (Hanover 1994)
15.114 R. Brennan, J. Christensen, W. Gruver, D. Kotak, D. Norrie, E. Leeuwen: Holonic manufacturing systems – a technical overwiew. In: *The Industrial Information Technology Handbook*, ed. by R. Zurawski (CRC Press, Boca Raton 2005) pp. 106/1–106/15
15.115 H. van Brussels, J. Wyns, P. Valckenaers, L. Bongaerts, P. Peeters: Reference architecture for holonic manufacturing systems - PROSA, Comput. Indust. **37**, 255–274 (1998)
15.116 S.M. Deen: *Agent Based Manufacturing – Advances in the Holonic Approach, Advanced Information Processing* (Springer, Berlin 2003)
15.117 A.-L. Barbasi: *Linked: The New Science of Networks* (Perseus, Cambridge 2002)
15.118 H. Kühnle, S.F. Schmelzer: The Individual in the Focus of the Factory, A New Paradigm – The Fractal View of the FACTORY. British Academy of Management – Annual Conference 1995 (British Academy of Management, Sheffield 1995) pp. 278–282
15.119 B.B. Mandelbrot: Self-affine fractals and fractal dimension, Phys. Scripta **32**, 257–260 (1985)
15.120 B.B. Mandelbrot, C.J.G. Evertsz: Multifractality on the harmonic measure on fractal aggregates, and extended self similarity, Physica A **177**, 386–393 (1991)
15.121 H.-J. Warnecke: *The Fractal Company – Revolution in Corporate Culture* (Springer, Berlin 1993)
15.122 E.N. Lorenz: The local structure of a chaotic attractor in four dimensions, Physica D **13**(1/2), 90–104 (1984)
15.123 H. Kühnle: L'entreprise fractale. In: *La modélization systémique en entreprise*, ed. by C. Braesch, A. Haurat (Pôle productique Rhône-Alpes, Paris 1995), pp. 263–272
15.124 K. Dooley: Complexity science models of organizational change. In: *Handbook of Organizational Change and Development*, ed. by S. Poole, A. van de Ven (Oxford Univ. Press, Cambridge 2004)
15.125 A. Lüder, J. Peschke, T. Sauter, S. Deter, D. Diep: Distributed intelligence for plant automation based on multi-agent systems – the PABADIS approach, J. Product. Plann. Control **15**(2), 201–212 (2004), Special Issue on Application of Multiagent Systems to PP&C
15.126 A. Klostermeyer, E. Klemm: Multi agent based architecture for plant automation. In: *The Industrial Information Technology Handbook*, Industrial Electronics Series, ed. by R. Zurawski (CRC Press, Boca Raton 2005) pp. 108/1–108/19
15.127 Wikipedia, the free encyclopedia: Cooperation, http://www.wikipedia.org/
15.128 R. Axelrod, W.D. Hamilton: The evolution of cooperation, Science **211**, 1390–1396 (1981)
15.129 A. Brandenburger, B.Nalebuff: Co-opetition, Interactive book, http://mayet.som.yale.edu/coopetition/
15.130 C.J. Clank: Strategic Alliances and Partnerships, Boyden web site, http://www.boyden.com/
15.131 B.R. Gomes-Casseres: *Alliances and Risk: Securing a Place in the Victory Parade* (Financial Times, May 9, 2000)
15.132 J.D. Bamford, B.R. Gomes-Casseres, S. Michael: *Mastering Alliance Strategy*, Jossey-Bass Business and Management Series (2003)
15.133 J. Child, D. Faulkner, S. Tallman: *Cooperative Strategy* (Oxford Univ. Press, New York 2005)

15.134 H.B. Thorelli: Networks: Between markets and hierarchies, Strategic Management J. **7**, 37–51 (1986)
15.135 Supply Chain Council: *Suplly-Chain Operations Reference Model, Version 7* (SCC, Pittsburgh 2004)
15.136 TNEE EU RTD Consortium: Green Book on Extended Enterprise (2003)
15.137 ARICON EU RTD Project: European Handbook for Virtual Enterprises (2005)
15.138 ALIVE EU RTD Project: Project Final Report (2002)
15.139 C.S. Snow, R.E. Miles, H.J. Coleman: Managing 21st century network organizations, Organ. Dyn. **20**, 5–20 (1992)
15.140 BIDSAVER IST EU RTD Project 10768: Final Report
15.141 R. Wigand, A. Picot, R. Reichwald: *Information, Organization and Management: Expanding Markets and Corporate Boundaries* (Wiley, Chichester 2004), reprinted
15.142 R.E. Kraut, R.S. Fish, R.W. Root, B.L. Chalfonte: *Informal Communication in Organizations: Form, Function, and Technology* (Sage, Newbury Park 1990)
15.143 C. Shannon, W. Weaver: *The Mathematical Theory of Communication* (Univ. of Illinois Press, Urbana 1963)
15.144 M. Bruhn: *Integrierte Unternehmens- und Markenkommunikation: Strategische Planung und operative Umsetzung*, 4th edn. (Schäffer-Poeschel, Stuttgart 2006)
15.145 P.A. Argenti: *Corporate Communication*, 3rd edn. (McGraw-Hill/Irwin, Boston 2003)
15.146 K. Miller: *Organizational Communication–Approaches and Processes*, 4th edn. (Wadsworth, Belmont 2005)
15.147 A.J. Zaremba: *Organizational Communication: Foundations for Business & Management*, 1st edn. (Thomson/South-Western, Mason 2003)
15.148 M. Poole (Eds.): *Human Resource Management: Critical Perspectives on Business and Management*, 3rd edn. (Routledge, London/New York 1999)
15.149 B. Hedley: Strategy and the business portfolio. In: *Strategic Marketing*, ed. by B.A. Weitz (Kent, Boston 1984) pp. 251–263
15.150 T.J. Gordon, H. Hayward: Initial experiments with the cross-impact matrix of forecasting, Futures **1**(2), 100–116 (1968)
15.151 W. Shen, D.H. Norrie, J.P. Barthès: *Multi-Agent Systems for Concurrent Intelligent Design and Manufacturing* (Taylor and Francis, London 2001)
15.152 T. Wagner, V.R. Lesser: Evolving real-time local agent control for large-scale multi-agent systems, Comput. Sci. **2333**, 51–68 (Springer, London 2001), archive, Revised Papers from the 8th International Workshop on Intelligent Agents VIII
15.153 R. Guimera, B. Uzzi, A. Spiro, A. Luis, N. Amaral: Team assembly mechanisms determine collaboration network structure and team performance, Science **308**(5722), 697–702 (2005)
15.154 A.-L. Barabasi: Network theory – the emergence of the creative enterprise, Science **308**(5722), 639–641 (2005)
15.155 H. Kühnle: Fractal extended enterprise: Framework and examples for multi-party supply chains. In: *Modern Industrial Engineering and Innovation in Enterprise Management, IEEM 2005*, Vol. 1, ed. by Y. Shuping, C. Xiaohui, Y. Yu (2005) pp. 211–217
15.156 A. Lomi, E.R. Larsen: *Dynamics of Organizations-Computational Modeling and Organization Theories* (The MIT Press, Cambridge 2001)
15.157 R. Dekkers: *(R)Evolution, Organizations and the Dynamics of the Environment* (Springer, New York 2005)
15.158 K. Miller: *Communication Theories: Perspectives, Processes, and Contexts*, 2nd edn. (McGraw-Hill, New York 2004)
15.159 K.-D. Thoben: *Kundenspezifische Produktion –y Prinzipien, Methoden und Werkzeuge, Habilitation Thesis* (Universitaet Bremen, Bremen 2000)
15.160 S.J. Childe: The extended enterprise – a concept of co-operation, Prod. Planning Control **9**(4), 320–327 (1998)
15.161 B. Kiesel, J. Klink: Die Renaissance der Kooperation, ZWF 93 **1–2**, 18–21 (1998), (in German)
15.162 G. Reinhart, S. Brandner: Produktdaten- und Prozeßmanagement in virtuellen Fabriken, Produktdatenmanagement **1**, 4–8 (2000), (in German)
15.163 S.L. Goldman, R.N. Nagel, K. Preiss: *Agile Competitors and Virtual Organizations: Strategies for Enriching the Customer* (Van Nostrand Reinhold, New York 1994)
15.164 S. Wurche: *Strategische Kooperationen – Theoretische Grundlagen und praktische Erfahrungen am Beispiel mittelständischer Pharmaunternehmen* (Dt. Univ.-Verl., Wiesbaden 1994), in German
15.165 P.B. Evans, T.S. Wurster: Strategy and the new economics of information, Harvard Bus. Rev. **75**(5), 70–93 (1997)
15.166 H. Håkansson, I. Snehotna: No business is an island: The network concept of business strategy, Scand. J. Manage. **4**(3), 256–270 (1989)
15.167 H. Håkansson, I. Snehotna (Eds.): *Developing Relationships in Business Networks* (Routledge, London 1995)
15.168 J.C. Anderson, H. Håkansson, J. Johanson: Dyadic relationships within a business network approach, J. Market. **58**, 0 (1994)
15.169 B. Axelsson: The development of network research – a question of mobilization and perseverance. In: *Industrial Networks. A New View on Reality*, ed. by B. Axelsson (Routledge, London 1992)
15.170 J.A. Carlisle, R.C. Parker: *Beyond Negotiation* (Wiley, Chichester 1989)
15.171 P. Smith Ring, A.H. van de Ven: Structuring cooperative relationships between organizations, Strategic Manage. J. **13**(7), 483–498 (1992)

15.172 R. Lamming: *Beyond Partnership: Strategies for Innovation and Lean Supply* (Prentice Hall, London 1993)
15.173 R.R. Kamath, J.K. Liker: A second look at japanese product development, Harvard Bus. Rev. **72**(6), 154 (1994)
15.174 C. Jones, W.S. Hesterly, S.P. Borgatti: A general theory of network goverance: Exchange conditions and social mechanisms, Acad. Manage. Rev. **22**(4), 911–945 (1997)
15.175 O.E. Williamson: *Markets and Hierarchies* (Free, New York 1975)
15.176 I. Hunt, H.S. Jagdev: Private discussions (2000)
15.177 E.K. Clemons, P.R. Kleindorfer: An economic analysis of interorganizational information technology, Decision Support Syst. **8**(5), 429 (1992)
15.178 M. Hardwick, D.L. Spooner, T. Rando, K.C. Morris: Sharing manufacturing information in virtual enterprises, Commun. ACM **39**(2), 46–54 (1996)
15.179 J. Browne, P.J. Sackett, J.C. Wortmann: Future manufacturing systems towards the extended enterprise, Comput. Ind. **25**(3), 235 (1995)
15.180 W.H. Davidow, M.S. Malone: *The Virtual Corporation: Structuring and Revitalizing the Corporation for the 21st Century* (HarperCollins, New York 1992)
15.181 B.R. Konsynski: Strategic control in the extended enterprise, IBM Syst. J. **32**(1), 111 (1993)
15.182 B. Gott: *Empowered Engineering for the Extended Enterprise – A Management Guide* (Cambashi, Cambridge 1996)
15.183 J. Browne, J. Harhen, J. Shivnan: *Production Management Systems – An Integrated Perspective*, 2nd edn. (Addison-Wesley, Boston 1996)
15.184 H.S. Jagdev, J. Browne: The extended enterprise: A context for manufacturing, J. Prod. Planning Control **9**(3), 216–229 (1998)
15.185 C. Thommessen: Network computing creating the new era of extended enterprise, Financial Times **5** (1996)
15.186 I. Hunter, D. Hassan, A. Gayoso, F. Garas: The eLSEwise vision, development routes and recommendations, Eng. Construct. Architect. Manage. **6**(1), 51–62 (1999)
15.187 Forbairt: *Virtual Corporation Defined, Summary Section for Forbairt Internet Report* (Forbairt, Dublin 1996)
15.188 C. Scholz: *Die virtuelle Organisation als Strukturkonzept der Zukunft?* (Univ. des Saarlandes, Saarbrücken 1997), http://www.orga.uni-sb.de/allgvo.html, in German
15.189 D. Skyrme : Networking to a Better Future Management Insights (1996), http://www.hiway.co.uk/skyrme/insights/insights.html, last accessed June 2005
15.190 IMS: http://www.img.org/ (Intelligent Manufacturing Systems, 1996), last accessed June 2006
15.191 C. Møller: Communication in inter-organisational operations towards a research framework in logistics and production management, APMS Int. Conf. Advances in Production Management Systems, IFIP Conference Proceedings, ed. by N. Okino, H. Tamura, S. Fujii (Chapman Hall, 1996)

Part C Complementary Material for Mechanical Engineers

16. Power Generation

Dwarkadas Kothari, P.M.V. Subbarao

The chapter contains 32 sections. Section 16.1 gives an introduction to the principle of energy supply. This section also provides the state of the art of the economics of various energy resources. Different types of fuels and their characteristics are discussed in Sect. 16.3. The conversion of different forms of energy has been explained in Sect. 16.5. Working principles of different power plants like gas turbines, the internal combustion (IC) engine, fuel cells, nuclear, and combined cycle system are discussed in Sects. 16.6–16.10.

Section 16.11 explores the inherent features of the integrated gasification combined cycle system. Various types of gasifiers and their working procedures are explained in this section. Section 16.12 provides updated information about magneto-hydrodynamic power generation and detailed information about various types of cogeneration system is also explained in Sect. 16.13.

Sections 16.14 and 16.15 explain the transformation of regenerative energies. These sections are mainly devoted to wind and solar energy conversion. Harvesting solar energy using solar ponds and solar chimneys is also explained in this section. The concept and working principle of the heat pump is explained in Sect. 16.16.

Section 16.17 contains the information about energy storage and distribution systems. Energy storage is used to offset the adverse effects of fluctuating demands for electricity and to assure a steady output from existing power plants. Various energy storage devices like pumped hydro, thermal energy, and hydrogen energy are described.

The furnace is the heart of a power generation system. Understanding its internal features and working principle is very important for a power plant professional. Section 16.18 satisfies these needs. It not only provides the characteristics of furnace combustion, but also provides the emission characteristics of furnace. Recent combustion technologies like fluidized bed combustion, bubbling fluidized bed combustion, and circulating fluidized bed combustion are also explored in Sect. 16.19.

Section 16.21 provides more details about the working principles of various types of burners. Inside the furnace the fuel must be evenly dispersed in the combustion airstream such that the fuel and air can come into intimate contact. Failure to achieve this results in unburnt or partially burnt fuel. The burner design attempts to achieve this by using a variety of techniques. Sections 16.22 and 16.23 facilitate understanding of various furnace accessories and technologies available to control emission.

The boiler is a key component in modern, coal-fired power plants; its concept, design, type, and integration into the overall plant considerably influence costs. The operating behavior and availability of the power plant are discussed in Sect. 16.24. Details of the various components of a steam generator are provided in Sect. 16.25.

Energy balance analysis and the efficiency calculation of furnace are described in Sects. 16.26–16.28. Thermodynamic calculations such as furnace design, boiler strength calculations, and heat transfer calculations are discussed in Sects. 16.29 and 16.30. Various types of nuclear reactors and their working principles are elaborated in Sect. 16.31. Finally, Sect. 16.32 is devoted to a discussion of future prospects and conclusions.

16.1 Principles of Energy Supply

Energy exists in many forms such as thermal energy, chemical energy, mechanical energy, potential energy, kinetic energy, and nuclear energy. Electrical energy is a desirable form of energy, because it can be generated centrally in bulk and transmitted economically over long distances. The requirement for energy is the demand for so many tonnes of coal, barrels of oil, cubic meters of gas, and so on. With the ever-increasing per-capita energy consumption and exponential growth in population, technologists already foresee the end of the Earth's non-replenishable fuel resources.

16.1.1 Planning and Investments

Investment planning for power plants requires a long-term plan, which covers facility investment such as the construction of new power plants or the replacement of existing plants with a newer one in an uncertain environments. Capital investment is a prerequisite for energy development as it is highly capital intensive. Investments in energy plants, equipment, and infrastructure (transportation, availability of fuel, water, communications, environment compatibility etc.) must be viewed

in the framework of economic growth, savings, and the size and degree of liberalization of capital markets.

According to the International Energy Agency (IEA), investment in electricity generation capacity will be about $4.6 trillion and installed capacity will rise from 3498 GW in 2000 to 7157 GW by 2030. As around 1000 GW of capacity is likely to be retired over this period, a total of 4700 GW of new build is required, costing around $4.28 trillion.

16.1.2 Economics of Gas

The yearly world demand for natural gas was 95.50 trillion cubic feet in 2003 and is rising gradually. In the year 2001 the consumption of natural gas was only 89.31 trillion cubic feet. Figure 16.1 provides information about the shares of electricity generation by fuel. Electricity generation from gas increased from 12.1% in 1973 to 19.4% in 2003.

Natural gas has the advantage over most other energy resources because of its great operational flexibility and the ease with which incremental gas supplies can be moved to generators. Various studies have revealed that gas is a preferred energy source for new generating capacity.

According to the statistics of the International Energy Agency, worldwide electricity production using gas in the year 2003 was 3 224 699 GW h out of a total electricity production of 16 741 884 GW h. According to the results of an IEA survey natural gas is tending to gain and coal to lose market share as the industry moves from a regulated position to a competitive environment.

Gas-fueled power plants have low capital cost. In 2003, 16% of the world's electricity was generated by natural gas. Technology transfer from developed countries will be required to meet this need. Pipeline transmission of gaseous fuel is capital intensive and allows less flexibility in the choice of buyers and sellers.

16.1.3 Economics of Electricity

A power plant should provide a reliable supply of electricity at minimum cost and minimum pollution to the consumer. The total price we pay for energy from power plants consists of:

1. Capital cost
2. Operating costs

Capital cost depends entirely on plant investment and includes:

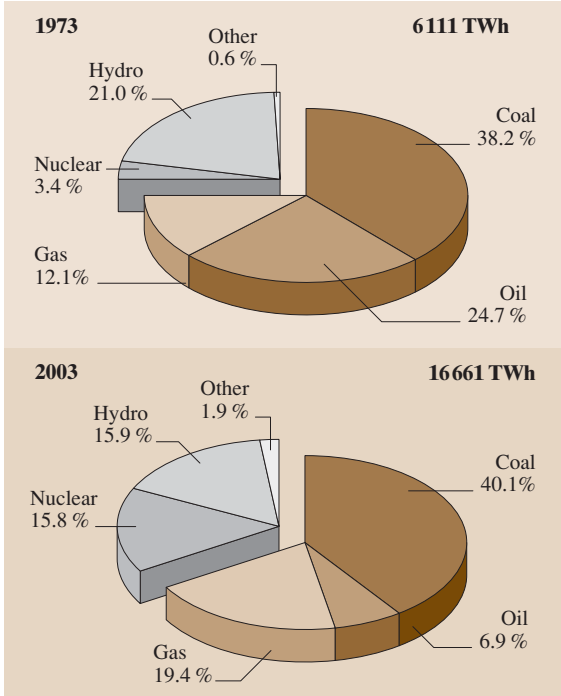


Fig. 16.1 Total energy shares of primary energy for electricity 2003 (Source: energy statistics from the IEA)

1. Interest
2. Depreciation
3. Taxes
4. Insurance

Operating cost includes the cost of operation, maintenance, and fuel.

Operating costs are inversely proportional to the capacity factor. Operating and maintenance costs mainly depend on the type of unit and the fuel used. Maintenance costs may be reduced if the unit is not operating or is operated at low load.

With the increase in economic growth, the consumption of electricity also increases. Electricity demand in India has been increasing rapidly, and the 534 billion kWh produced in 2002 was almost double the 1990 output, although it represents only 505 kWh per capita for the year. The amount and cost of electricity depends upon the fuel used.

16.1.4 Economics of Remote Heating

The cost of generating useful heat should be identified with the value of the *lost* electricity. The total

cost of waste heat has two components: the production cost at the plant and the distribution cost. The distribution component is calculated from the capital charges and maintenance expense for the pipeline system conveying warm water or steam from the plant to customers.

District heating is already widely used in central core areas of large cities. Many large buildings can be heated by steam or hot water from a single large central combustion plant. Also, space cooling can be provided by using the hot water or steam to actuate an absorption-type refrigeration plant.

16.2 Primary Energies

Primary energy is contained in raw fuels and any other forms of energy received by a system as input to the system. Primary energy is transformed in energy conversion processes to more convenient forms of energy and cleaner fuels. The most important primary energy sources are the carbon-based fossil energy sources. Fossil fuels (oil, coal, and natural gas) are called nonrenewable energies, and come from the long-term decomposition of plant and animal matter over millions of years. Sun is the main source of energy from which all of the above energy resources are derived.

16.3 Fuels

Fuels are chemical substances which may be burned in oxygen to generate heat. They mainly consist of carbon and hydrogen and sometimes a small amount of sulfur or minerals, and may be solid, liquid, or gaseous. Coal and coke are examples of solid fuels. Petroleum oils are usually a mixture of several liquid fuels. Gaseous fuels may be a mixture of gases such as methane (CH_4), ethane (C_2H_6) and so on.

16.3.1 Solid Fuels

Solid fuel is a term given to various types of solid materials that provide energy. This energy is usually released by combustion. Coal and coke are examples of solid fuels.

16.3.2 Liquid Fuels

Most liquid fuels are derived from fossil fuels. These can be classified according to their volatility (the ease with which they evaporate and turn into vapor). The most volatile fuels are gasoline and kerosene. Less volatile fuels are used in diesel engines and residual fuels, of varying viscosities, are often used in boilers. Ethanol produced from the fermentation of sugar is a prominent liquid fuel.

16.3.3 Gaseous Fuels

Gas is a preferred fuel, the combustion of which offers more environmental friendliness over the other fossil fuels. It burns more readily and completely than other fuels. Gaseous fuels are the most convenient, requiring the least amount of handling, and are the most maintenance free. Gas is odorless and colorless. Because gaseous fuels are in a molecular form, they are easily mixed with the air as required for combustion, and are oxidized in less time than is required to burn other types of fuel. A mixture of methane (CH_4) and ethane (C_2H_6) is an example of a gaseous fuel.

16.3.4 Nuclear Fuels

Fuels such as uranium or thorium that can be used in nuclear reactors as a source of electricity are called nuclear fuels. The energy derived during fission or fusion processes is called nuclear energy. Examples of nuclear fuels are: ^{235}U, ^{238}U, and ^{239}Pu.

16.3.5 Regenerative Energies

Regenerative or renewable energies are those energy sources or energy carriers that naturally renew themselves within human timescales. Regenerative energies

are available everywhere. Effective utilization of these resources is a very challenging task [16.1]. The renewable sources of energy include hydropower, solar, wind, and biomass [16.2].

16.4 Transformation of Primary Energy into Useful Energy

Energy has become an essential driving force of the economy. In fact, it assumes numerous forms: chemical energy in fossil fuels or the biomass, kinetic energy in waterfalls or the wind, electromagnetic energy from the sun, nuclear energy in uranium, as well as the electrical or thermal energy that is put to numerous uses. The sun is the major source of energy from which all energy resources are derived. Figure 16.2 shows the various forms of energy derived from the sun. It represents all the fossil fuels (oil, gas, and coal) derived from the sun by various means of transformation such as vegetation, chemical energy conversion, and finally fossilization.

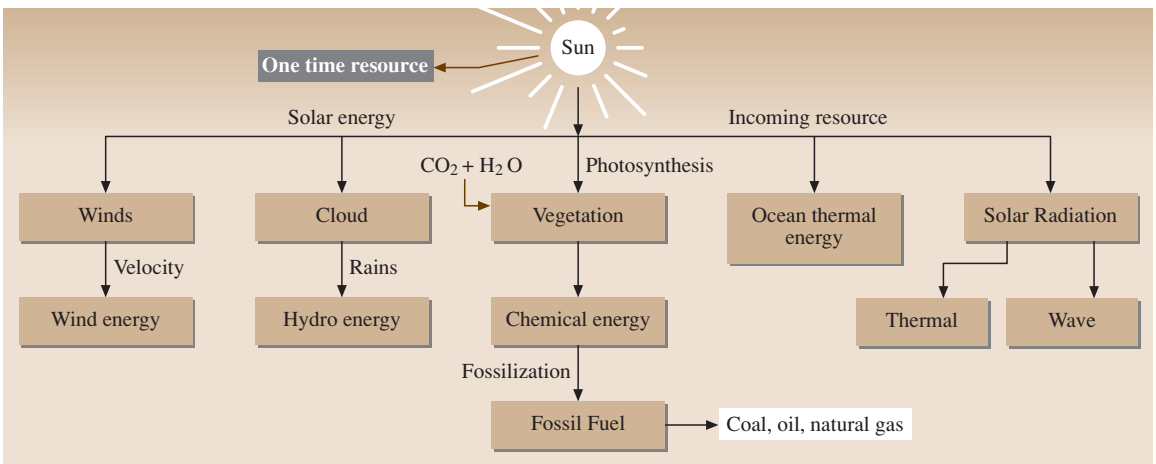


Fig. 16.2 Energy system diagram

16.5 Various Energy Systems and Their Conversion

Conversion of one form of energy into another form of energy is essential in order to utilize the maximum potential of energy resources. Efficient generation of electricity from various energy resources is one of the challenging tasks for the scientist. Efficient utilization of various sources of energy leads to a strong economy and promotes the wealth of the nation. Figures 16.3–16.5 depict various energy systems and their conversion process.

The sun heats our atmosphere unevenly, so some patches become warmer than others. These warm patches of air rise and other air blows in to replace them – and we feel a wind blowing. We can use the energy in the wind by building a tall tower, with a large propeller on the top.

The current global average conversion efficiency for coal-fired electricity generation is 34% and that for gas-fired electricity generation 37%. Modern coal-fired power plants are operating at a higher efficiency of 43–48 %.

16.5.1 Generation of Electrical Energy

Electric energy is the flow of electric power or charge. It is a secondary energy source, which means that we get it from the conversion of other sources of energy, such as coal, natural gas, oil, nuclear power, and other natural sources, which are called primary sources. The energy sources we use to make electricity can be renewable or nonrenewable, but electricity itself is either renew-

able or nonrenewable. Electricity generation from the combustion of fuel is reported under combustion-based power station.

16.5.2 Steam Power Cycle

The Rankine cycle is a thermodynamic cycle and is used in a variety of power plants. The simplest arrangement of the steam power plant is that without regeneration and reheat, as shown in Fig. 16.7. A simple Rankine cycle consists of four main components (steam generator, turbine, condenser, and pump). Additional components are sometimes added to enhance cycle performance and to improve efficiency. This cycle is named after William John Macquorn Rankine (1820–1872), who established it as the fundamental cycle for a steam power plant. In a simple steam power plant, which works on a Rankine cycle, heat is added reversibly at a constant pressure. The efficiency of the Rankine cycle is a function of the temperature of heat rejection and the mean temperature of heat addition. The higher the mean temperature of heat ad-

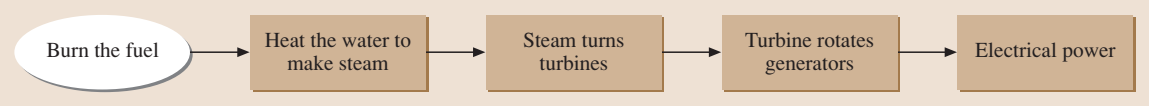


Fig. 16.3 Structure of energy systems – fossil sources

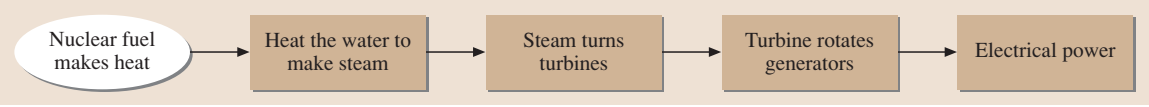


Fig. 16.4 Structure of energy system – nuclear fuels

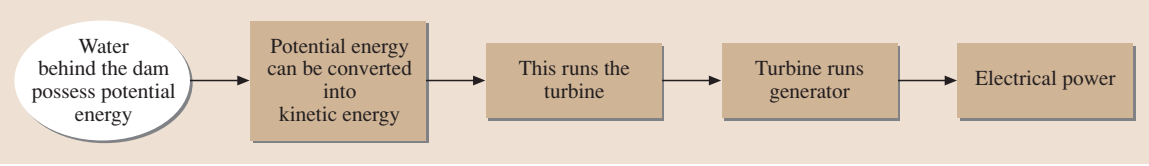


Fig. 16.5 Structure of energy systems – hydro fuels

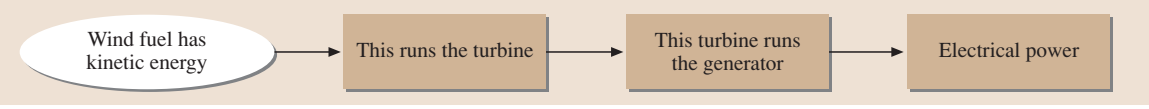


Fig. 16.6 Structure of energy systems – wind fuels

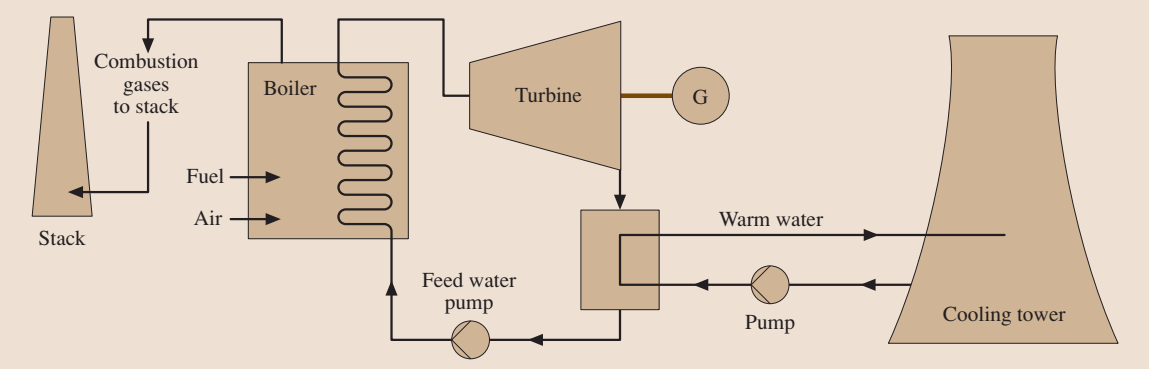


Fig. 16.7 Layout of a simple Rankine cycle power plant

dition the higher the efficiency. The working fluid in a Rankine cycle follows a closed loop and is reused constantly.

16.5.3 Process of the Rankine Cycle

There are four processes in the Rankine cycle, each changing the state of the working fluid. These states are identified in the T–S (temperature–entropy) diagram shown in Fig. 16.8.

Process 1–2. First the working fluid is pumped (isentropically) from low to high pressure by a pump. Pumping requires a power input (for example, mechanical or electrical).

Process 2–3. The high-pressure liquid enters a boiler, where it is heated at a constant pressure by an external heat source to become a superheated vapor. The common heat sources for power plant systems are coal, natural gas, and nuclear power.

Process 3–4. The superheated vapor expands through a turbine to generate power output; ideally this expansion is isentropic. This decreases the temperature and pressure of the vapor.

Process 4–1. The vapor then enters a condenser, where it is cooled to become a saturated liquid. This liquid then reenters the pump and the cycle repeats.

In order to improve coal-fired power plant efficiency, leading to a proportional reduction in coal consumption and carbon dioxide emissions, it is widely accepted that the domestic power industry must move from subcritical to supercritical steam cycles.

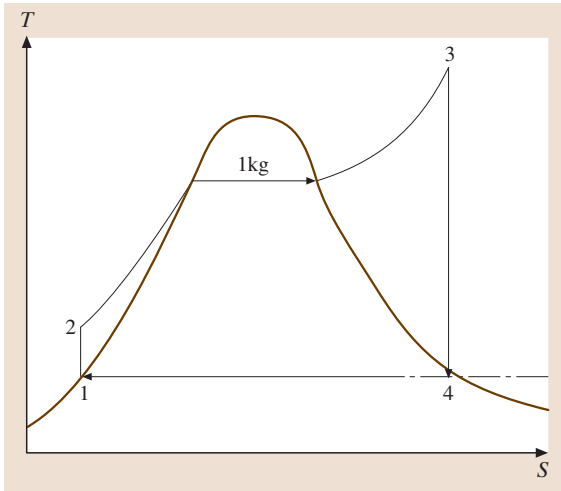


Fig. 16.8 T–S diagram of a simple Rankine cycle

Reaching 45% efficiency is possible with the help of a supercritical (SC) Rankine steam cycle employing a reheat and regeneration mode and operating at 250 bar and 540 °C. Today, ultra-supercritical steam (USC) parameters of 300 bars and 600/600 °C can be realized, resulting in 42% (HHV – higher heating value) efficiency for bituminous-coal-fired power plants [16.3].

The improved efficiency represents reductions of about 15% in all emissions including CO_2, compared to those from installed capacity. The challenges of coal-based power generation are environmental; the future technologies are near-zero emission and high-efficiency plants equipped to reduce CO_2 emission.

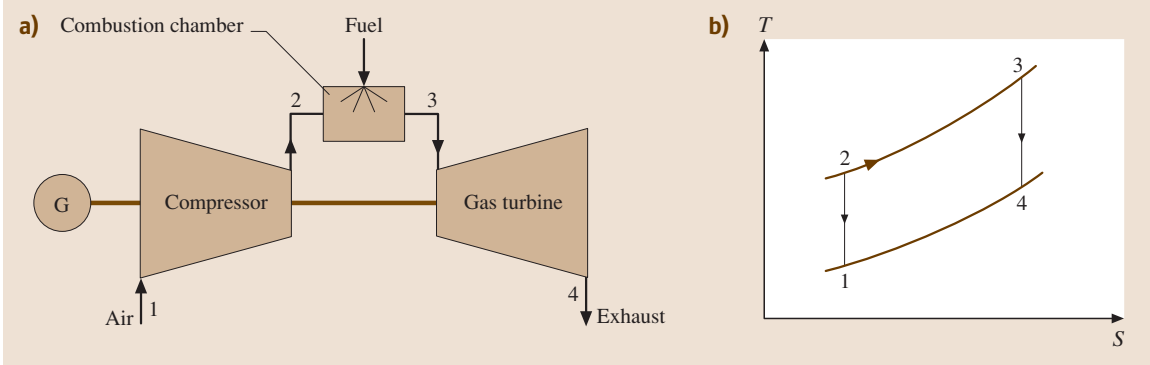


Fig. 16.9 Arrangement of an open-cycle gas turbine plant

16.6 Direct Combustion System

The economics of power generation by gas turbines is now quite attractive due to their low capital cost and high reliability and flexibility in operation. Another outstanding feature is their capability to start quickly and to use a wide variety of fuels from natural gas to residual oil or powdered coal.

16.6.1 Open-Cycle Gas Turbine Power Plant

The essential components of a gas turbine power plant are the compressor, combustion chamber, and the turbine. The air standard cycle of gas turbine power plant is the Brayton cycle shown in Fig. 16.9. It consists of two reversible adiabatic processes and two constant-pressure processes. Gas turbine plants can be operated either in an open or closed system configuration.

Analysis

1–2: Work input

$$w_{\text{comp}} = h_2 - h_1 = c_{\text{p}}(T_2 - T_1)\,, \tag{16.1}$$

2–3: Heat input

$$q_{\text{in}} = h_3 - h_2 = c_{\text{p}}(T_3 - T_2)\,, \tag{16.2}$$

3–4: Work output

$$w_{\text{tur}} = h_3 - h_4 = c_{\text{p}}(T_3 - T_4)\,, \tag{16.3}$$

4–1: Heat rejection

$$q_{\text{out}} = h_4 - h_1 = c_{\text{p}}(T_4 - T_1)\,. \tag{16.4}$$

Isentropic processes

$$\frac{p_2}{p_1} = \left(\frac{v_1}{v_2}\right)^{\gamma} = \left(\frac{T_2}{T_1}\right)^{\frac{\gamma}{\gamma-1}}\,, \tag{16.5}$$

$$\frac{p_3}{p_4} = \left(\frac{v_4}{v_3}\right)^{\gamma} = \left(\frac{T_3}{T_4}\right)^{\frac{\gamma}{\gamma-1}}\,. \tag{16.6}$$

Constant-pressure processes

$$p_3 = p_2 \quad \text{and} \quad p_4 = p_1\,; \tag{16.7}$$

$$r_{\text{p}} = \frac{p_2}{p_1} = \frac{p_3}{p_4} = \left(\frac{v_1}{v_2}\right)^{\gamma} = \left(\frac{v_4}{v_3}\right)^{\gamma} = \left(\frac{T_2}{T_1}\right)^{\frac{\gamma}{\gamma-1}} = \left(\frac{T_3}{T_4}\right)^{\frac{\gamma}{\gamma-1}}\,, \tag{16.8}$$

where r_{p} is the pressure ratio

$$T_2 = T_1(r_{\text{p}})^{\frac{\gamma-1}{\gamma}} = T_1\rho\,, \tag{16.9}$$

and

$$\rho = (r_{\text{p}})^{\frac{\gamma-1}{\gamma}}\,, \tag{16.10}$$

$$T_4 = \frac{T_3}{(r_{\text{p}})^{\frac{\gamma-1}{\gamma}}} = \frac{T_3}{\rho}\,, \tag{16.11}$$

$$\eta_{\text{th}} = \frac{w_{\text{net}}}{q_{\text{in}}} = \frac{c_{\text{p}}\left(\frac{T_3}{\rho} - T_1\right)}{c_{\text{p}}\,(T_3 - \rho T_1)} = \frac{1}{r_{\text{p}}^{\frac{\gamma}{\gamma-1}}}\,, \tag{16.12}$$

$$w_{\text{net}} = c_{\text{p}}\left[T_3\left(\frac{\rho-1}{\rho}\right) - T_1(\rho-1)\right]\,, \tag{16.13}$$

$$= c_{\text{p}}(\rho-1)\left(\frac{T_3}{\rho} - T_1\right)$$

$$= c_{\text{p}}\left(\frac{\rho-1}{\rho}\right)(T_3 - \rho T_1)\,. \tag{16.14}$$

The thermal efficiency can also be written as

$$\eta_{\text{th}} = \frac{1}{\rho} = \frac{1}{r_{\text{p}}^{\frac{\gamma}{\gamma-1}}}\,. \tag{16.15}$$

It may be noted that in a simple gas turbine cycle the cycle efficiency is a function of the pressure ratio only. The gas turbine inlet temperature is an important parameter of efficiency. The present state of the art temperature is 1570 K, but research on closed-cycle steam cooling of turbine blades, protective surface coating of combustor liners, and new ceramic structural parts of the turbine are areas of research that will lead to higher gas turbine inlet temperatures.

Merits and Demerits of the Brayton Cycle

1. Very compact, which is why it is used in aircraft.
2. It demands extremely high quality and costlier fuel.
3. The pressure of the exit gases should always be just above atmospheric pressure.
4. The compressor requires a large power input. In consumes more power than is produced from the steam turbine.
5. It has a lower cycle efficiency, due to the large exhaust loss.

16.7 Internal Combustion Engines

The internal combustion (IC) engine is one of the greatest inventions today. Internal combustion engines can be run by p.t.o., diesel, gasoline, methane, or natural gas. Internal combustion engines for power production are generally fueled by diesel.

Reciprocating engines have usually been employed for distributed power generation for the past few decades. Power generation from a diesel engine generator is the most cost-competitive technology to provide power to a small number of consumers. It is appropriate for an electrical load of about 0.01–50 MW.

Internal combustion engines have one or more cylinders in which the combustion of fuel takes place. The engine, which is connected to the shaft of the generator, provides the mechanical energy to drive the generator to produce electricity. The following are the benefits of IC engines for power production:

1. Easy to transport and install.
2. Like gas turbines, they are usually operated during periods of high demand for electricity.

16.8 Fuel Cells

A fuel cell is an electrochemical device that converts the chemical energy of the fuel directly into electrical energy. It consists of an electrolyte layer in contact with a porous anode and cathode on the sides. In a typical fuel cell gaseous fuels are fed continuously to the negative electrode and an oxidant is fed continuously to the positive electrode.

Electrochemical reactions take place at the electrodes to produce an electric current. In general, the oxidation of H_2, CO, CH_4, and higher hydrocarbons in fuel cells that produces power also produces rejected heat. This heat arises from two sources: when the entropy decreases, ΔS resulting from the overall oxidation reaction, accompanying the usual decrease in the number of moles of gases from reactants to products and due to the irreversible process occurring in the operation of the cell. Heat from these two sources must be rejected from the fuel cell in order to maintain the temperature at a desired level. In the case of hybridization, the efficiency is the ratio of the sum of stack electricity and the power generated by the bottoming cycle to the lower calorific value of the fuel consumed. This can be accomplished by three cycles, namely the regenerative Brayton cycle, the combined Brayton–Rankine cycle, and the steam-turbine-operated Rankine cycle.

The regenerative Brayton cycle in Fig. 16.10 shows a gas turbine compressor for the air flow to the cell. The flow then passes through a countercurrent recuperative heat exchanger to recover heat from the combustion product gases leaving the gas turbine. The air and the fuel streams then pass into the cathode and anode compartments of the fuel cell(s). The air and fuel streams

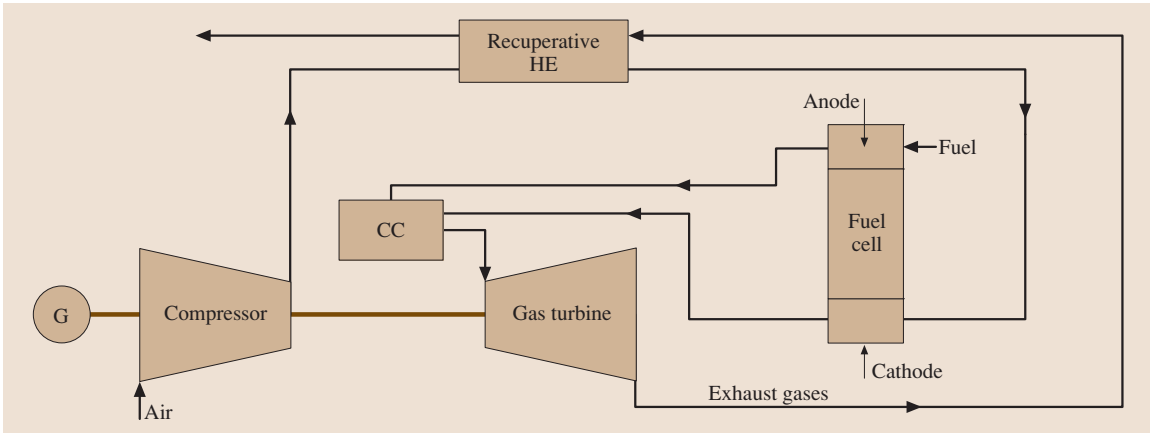


Fig. 16.10 Regenerative Brayton-cycle fuel-cell power system. (CC – combustion exchanger, HE – heat exchanger)

leaving the cell(s) enter the combustor, where they mix and the residual fuel burns. The combustion products enter the turbine, expand, and generate additional power. The turbine exhaust gases pass through the recuperative exchanger to the stack.

The overall solid oxide fuel cell reactions are

$$H_2 + \tfrac{1}{2}O_2 \rightarrow H_2O\,, \qquad \Delta H_{298} = -241\,\mathrm{kJ/mol}\,, \tag{16.16}$$

$$CO + \tfrac{1}{2}O_2 \rightarrow CO_2\,, \qquad \Delta H_{298} = -283\,\mathrm{kJ/mol}\,. \tag{16.17}$$

The ideal or equilibrium potential (E) for the above overall reactions can be calculated by the Nernst equation

$$E = E^0 + \frac{RT}{2F}\ln\left(\frac{P_{H_2}P_{O_2}^{1/2}}{P_{H_2O}}\right) \quad \text{and}$$

$$E = E^0 + \frac{RT}{2F}\ln\left(\frac{P_{CO}P_{O_2}^{1/2}}{P_{CO_2}}\right)\,. \tag{16.18}$$

Fuel cell power output

$$= \text{voltage (cell potential)} \times \text{current (load)}\,. \tag{16.19}$$

The total power and total efficiencies of the hybrid cycle are calculated as

$$P_{SOFC} = N_{cell}P_{cell}\eta_{DC/AC}\,, \tag{16.20}$$

$$P_{GT} = (P_{EXP} - P_{COMP})\eta_{GEN}\,, \tag{16.21}$$

$$P_{TOTAL} = P_{SOFC} + P_{GT} - P_{AUX}\,, \tag{16.22}$$

$$\eta_{ele} = \frac{P_{TOTAL}}{\sum N_c \mathrm{LHV}_c}\,, \tag{16.23}$$

where P_{SOFC} is the electrical power output from the fuel cell, P_{GT} is the power output from the gas turbine, P_{aux} is the power consumed by the auxiliary units, LHV_c is the lower heating value, and N_c is the number of cells.

The Siemens Westinghouse Power Corporation of Pittsburgh developed and fabricated the first advanced power plant to combine a solid oxide fuel cell and a gas turbine. The microturbine generator was manufactured by Northern Research and Engineering Corporation of Woburn. The factory acceptance test was completed in April 2000. Southern California Edison is operating the new hybrid plant at the National Fuel Cell Research Center at the University of California, Irvine. A year of testing in a commercial setting will be performed at this site. The system cycle is expected to generate electric power at 55% efficiency. The pressurized system will generate 220 kW of power and be operated at a pressure of 3 atm. The fuel cell is made up of 1152 individual tubular ceramic cells and generates about 200 kW of electricity. The microturbine generator will produce an additional 20 kW of electricity at full power. No sulfur dioxide pollutants will be released into the air. Nitrogen oxide emissions are likely to be less than 1 ppm. A 320 kW hybrid system is also in the planning stage. An initial commercial offering of a 1 MW fuel-cell microturbine power plant in late 2002 were the end results of this Department of Energy/Siemens Westinghouse partnership program. Global electricity generating capacity from fuel cells will grow from just 75 MW in 2001 to 15 000 MW by 2010. The cost of generating electricity (> $2000/kW) is very high for fuel cell, which restricts its wideusage.

16.9 Nuclear Power Stations

Due to the depleting nature of coal reserves, switching from coal to nuclear is mandatory today. In fact nuclear fuels have tremendous potential to release vast amount of energy from fission reactions. Uranium is one of the most important nuclear fuels used in nuclear power plants. When uranium is bombarded with neutrons, fission reactions take place, releasing neutrons and tremendous amounts of energy. In 1942 Enrico Fermi used uranium to produce the first controlled chain reaction. The fission of uranium atoms in the reactor imparts the heat to produce steam for the generation

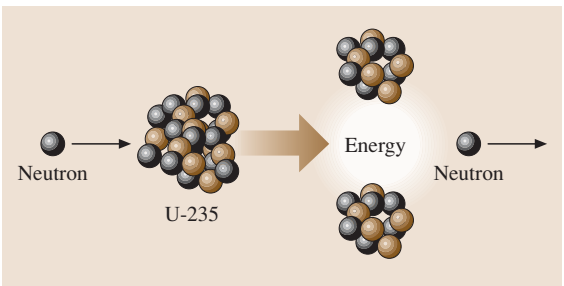


Fig. 16.11 Fission of a uranium 235 nucleus

electricity. When one pound of pure uranium is completely fissioned, it creates as much energy as burning of 1500 short tons of coal.

16.9.1 Basic Principles of Nuclear Energy

The powerful theory used to estimate the amount of energy released in a fission process is Albert Einstein's theory of relativity. Einstein's theory of relativity states that,

$$E = mc^2 , \tag{16.24}$$

where E is the energy in a substance, which equals the mass (m) of that substance multiplied by the speed of light squared (c^2). Nowadays most nuclear reactors are of fission type. In nuclear fission, the nuclei of atoms are split, causing energy to be released. The element uranium is the main fuel used in nuclear fission to produce energy since it has many useful properties: uranium nuclei can easily be split by shooting neutrons at them, and multiple neutrons released during this process are used to split other uranium, nuclei as shown in Fig. 16.11. This phenomenon is known as a chain reaction.

16.9.2 Types of Nuclear Power Plants

Nuclear power plants are classified according to the type of reactor used.

Fission Reactors

Fission power reactors generate heat by the nuclear fission of fissile isotopes of uranium and plutonium.

Various types of nuclear reactors are used in practice for power plants, which may be categorized into three classes:

- *Thermal reactors* use a neutron moderator to slow down or *moderate* the rate of production of fast neutrons by fission, to increase the probability that they will produce fission and thus sustain the chain reaction.
- *Fast reactors* sustain the chain reaction without the need for a neutron moderator.
- *Subcritical reactors* use an outside source of neutrons.
- Light-water reactor (LWR):
  - Pressurized-water reactor (PWR)
  - Boiling-water reactor (BWR)
- Graphite-moderated reactor:
  - Magnox
  - Advanced gas-cooled reactor (AGR)
  - Chernobyl type
  - Pebble-bed reactor (PBMR)
- Heavy-water moderated reactor:
  - CANDU (CANadian Deuterium Uranium)

16.10 Combined Power Station

When two cycles are combined, the cycle operated at the higher temperature is called the *topping cycle*. The waste heat it produces is then used in a second process that operates at a lower temperature level and is therefore called the *bottoming cycle* [16.4]. A combined cycle system is a sandwich of two different cycles,

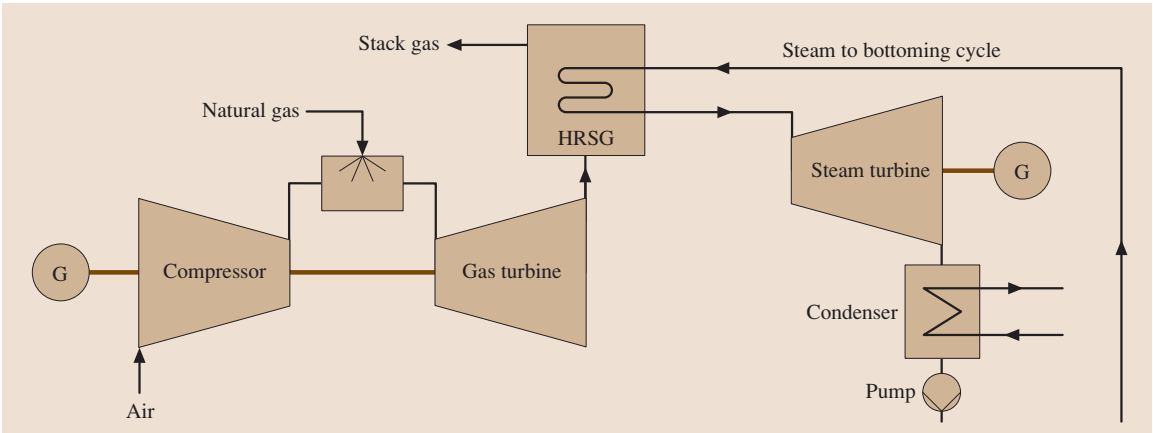


Fig. 16.12 Schematic diagram of a combined-cycle power plant

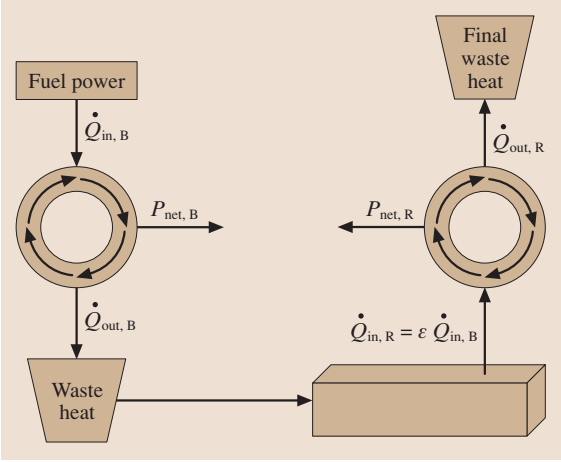


Fig. 16.13 Analysis of the combination of two different cycles

normally gas turbine and steam turbine cycles. The combination of two different cycles with different working media is interesting because the advantages of one complement the other [16.5].

The simplest form of combined cycle power plant, as shown in Fig. 16.12, is a single-pressure system. The major components present in the cycle is a single-pressure heat recovery steam generator (HRSG), a steam turbine, a water-cooled condenser, and a deaerater. The air is compressed inside the compressor and is supplied to the combustion chamber. The fuel (natural gas) from the main supply line is compressed in gas compressors and injected into the combustion chamber. The air and fuel mix together and then the combustion products leave the combustion chamber.

The combustion products from the combustion chamber enter the gas turbine and there expand to a low pressure and hence power is produced. The exhaust gas leaving the gas turbine is then supplied to the heat-recovery steam generator.

In the HRSG water flows in the direction opposite to the exhaust gas leaving the gas turbine and hence heat transfer takes place between the two fluids. The water absorbs heat from the exhaust gas, turns into steam, is superheated, and is then supplied to the steam turbine.

16.10.1 Thermodynamic Analysis of the Combined Cycle System

From Fig. 16.13 the net power output of the Brayton cycle is

$$P_{\text{net,B}} = \eta_{\text{B}}\,\dot{Q}_{\text{in,B}}\,, \tag{16.25}$$

where η_{B} is the Brayton efficiency.

The rate of heat rejected in the Brayton cycle is

$$\dot{Q}_{\text{out,B}} = (1-\eta_{\text{B}})\dot{Q}_{\text{in,B}}\,. \tag{16.26}$$

The rate of heat input to the Rankine cycle is

$$\dot{Q}_{\text{in,R}} = \varepsilon\,\dot{Q}_{\text{out,B}} = \varepsilon(1-\eta_{\text{B}})\dot{Q}_{\text{in,B}}\,, \tag{16.27}$$

where ε is the effectiveness.

The net power output of the Rankine cycle is

$$P_{\text{net,R}} = \eta_{\text{R}}\,\dot{Q}_{\text{in,R}} = \eta_{\text{R}}\varepsilon(1-\eta_{\text{B}})\dot{Q}_{\text{in,B}}\,, \tag{16.28}$$

where η_{R} is the Rankine efficiency.

The net power output of the combined cycle is

$$P_{\text{tot}} = P_{\text{net,B}} + P_{\text{net,R}}\,, \tag{16.29}$$

$$P_{\text{tot}} = \eta_{\text{B}}\,\dot{Q}_{\text{in,B}} + \eta_{\text{R}}\varepsilon(1-\eta_{\text{B}})\dot{Q}_{\text{in,B}}\,. \tag{16.30}$$

Overall Efficiency of the Combined Cycle Power Plant

The overall efficiency is defined as the ratio between the total power output to the heat added to the system

$$\eta_{\text{ov}} = \frac{P_{\text{tot}}}{\dot{Q}_{\text{in,B}}} \tag{16.31}$$

$$\eta_{\text{ov}} = \frac{\eta_{\text{B}}\,\dot{Q}_{\text{in,B}} + \eta_{\text{R}}\varepsilon(1-\eta_{\text{B}})\dot{Q}_{\text{in,B}}}{\dot{Q}_{\text{in,B}}} \tag{16.32}$$

$$\eta_{\text{ov}} = \eta_{\text{B}} + \eta_{\text{R}}\varepsilon(1-\eta_{\text{B}})\,. \tag{16.33}$$

16.11 Integrated Gasification Combined Cycle (IGCC) System

16.11.1 Introduction

IGCC technology is a promising technology which includes the benefit of gasification. A variety of fuel such as bituminous coal, sub-bituminous coal, lignite, petroleum coke, heavy oil, orimulsion, and biomass can provide the input to the IGCC cycle. Figure 16.14 shows the layout of integrated gasification combined cycle. Any grade of coal is gasified under pressure in the gasifier.

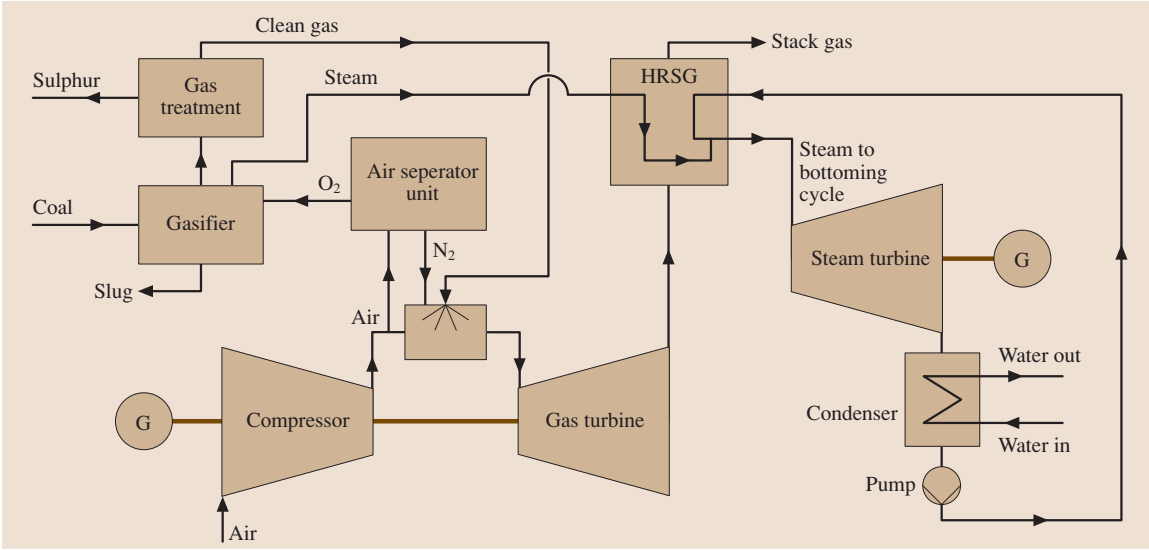


Fig. 16.14 Layout of an integrated gasification combined-cycle power plant

Syngas from the gasifier is cleaned of its hydrogen sulfide, ammonia, and particulate matter and is fed into a gas turbine where it is mixed inside the combustion chamber with hot pressurized air from the compressor. The final hot combustion products drive the gas turbine. The hot combustion gases from the gas turbine are used to produce steam in the steam generator. The steam drives the steam turbine, which produces 30–40% of the total electricity output. An air separation unit is also employed in modern IGCC power plants. Nitrogen and oxygen are completely separated in the air separator, from which pure oxygen is fed into the gasifier to reduce carbon dioxide emissions and the inert gas nitrogen is very well utilized in the gas turbine.

The technology makes use of the thermodynamic advantages provided by combining two different cycles: a gas turbine cycle and steam turbine cycle. Cleaning the gas before combustion provides benefits over the treatment of flue gases: a much smaller quantity of gas has to be treated and in addition the composition of the coal gas is such that it allows easier purification. The purification process could possibly be extended and could also permit the elimination of exhaust carbon dioxide. Hence, this technology has been proposed as the basis for a low-CO_2-emission coal power plant with CO_2 capture. Ample gasification processes have to be selected and the integration and optimization of all the processes are crucial for the overall efficiency. Coal gas clean up is another critical issue.

16.11.2 Environmental Benefits

1. Coal that contains a higher sulfur content can be very well utilized in IGCC plant. During the coal gasification process the sulfur in the coal appears as hydrogen sulphide; capturing hydrogen sulphide is not a tedious task. In some IGCC plants the sulfur can be extracted in a form that can be sold commercially.
2. Likewise, nitrogen typically exits as ammonia and can be scrubbed from the coal gas by processes that produce fertilizers or other ammonia-based chemicals.
3. If oxygen is used in a coal gasifier instead of air, carbon dioxide is emitted as a concentrated gas stream. In this form, it can be captured more easily and at lower costs for ultimate disposition in various sequestration approaches.

16.11.3 Efficiency Benefits

Efficiency gain is another benefit of IGCC plants. The fuel efficiency of IGCC power plant can be boosted to 50% or more. Future insights that integrate a fuel cell could achieve even higher efficiencies, maybe in the 60% range, which is nearly twice the value of today's typical coal-fired power plants. Higher efficiencies translate into more economical electric power and potential savings for ratepayers. A more efficient plant also uses less fuel to generate power, meaning that

less carbon dioxide is produced. IGCC plants with the flexibility to produce chemicals such as ammonia and hydrogen along with electricity make this a promising technology for future generations.

16.11.4 The Science of Coal Gasification

Coal gasification involves the chemical reaction of coal, steam, and air or oxygen at high temperatures to produce a mixture of hydrocarbon gases, typically carbon monoxide, carbon dioxide, and methane as well as hydrogen sulfide.

16.11.5 Chemical Reactions

Coal combustion, which is the exothermic reaction of coal with oxygen or air to produce carbon dioxide and water, is a fundamental part of coal gasification, using 20–40% of the oxygen or air required for complete combustion.

The purpose of this partial combustion is to supply the energy necessary to complete the gasification of the coal particles.

$$\mathrm{C(s)} + \tfrac{1}{2}\mathrm{O_2} \rightarrow \mathrm{CO\ (partial\ combustion)}\,,$$
$$\Delta H_\mathrm{R} = -1110.5\,\mathrm{MJ/kmol}\,, \qquad (16.34)$$

where ΔH_R is the standard heat of reaction at 298 K and atmospheric pressure

This partial combustion reaction is exothermic, that is, it liberates heat, as signified by the negative sign.

Furthermore the reaction of carbon does not stop at CO_2, but any remaining oxygen rapidly reacts with CO in the gas phase to produce CO_2

$$\mathrm{CO} + \tfrac{1}{2}\mathrm{O_2} \rightarrow \mathrm{CO_2}\,,$$
$$\Delta H_\mathrm{R} = -283.1\,\mathrm{MJ/kmol}\,. \qquad (16.35)$$

The solid carbonaceous material that is not combusted by oxygen reacts endothermically with carbon dioxide, hydrogen, and methane

$$\mathrm{C} + \mathrm{CO_2} \rightarrow 2\mathrm{CO}\,,$$
$$\Delta H_\mathrm{R} = +172.0\,\mathrm{MJ/kmol}\,. \qquad (16.36)$$

This reaction is called the Boudouard reaction.

In order to control the high temperatures resulting from the C(s)-O_2 reactions, and to increase the heating value of the product gas through the production of hydrogen, steam is often added as a reactant.

$$\mathrm{C(s)} + \mathrm{H_2O(g)} \rightarrow \mathrm{CO} + \mathrm{H_2}\,,$$
$$\Delta H_\mathrm{R} = +131.4\,\mathrm{MJ/kmol}\,. \qquad (16.37)$$

In a coal–steam or oxygen–steam, the homogeneous water-gas shift reaction is also important:

$$\mathrm{CO} + \mathrm{H_2O(g)} \rightarrow \mathrm{CO_2} + \mathrm{H_2}\,,$$
$$\Delta H_\mathrm{R} = -41.0\,\mathrm{MJ/kmol}\,. \qquad (16.38)$$

Hydrogen is added as a reactant in order to increase the quantity of methane. Water gas from which the CO_2 has been removed is called *synthesis gas*. Synthesis gas can also be used to produce methane, or synthetic natural gas (SNG)

$$\mathrm{CO(g)} + 3\mathrm{H_2(g)} \rightarrow \mathrm{CH_4(g)} + \mathrm{H_2O(g)}\,, \qquad (16.39)$$
$$2\mathrm{CO(g)} + 2\mathrm{H_2(g)} \rightarrow \mathrm{CH_4(g)} + \mathrm{CO_2(g)}\,. \qquad (16.40)$$

In fuel-rich combustion, the sulfur in the coal is released mainly as hydrogen sulfide with a small amount of carbonyl sulfide and the fuel-bound nitrogen is released as elemental nitrogen, ammonia, and hydrogen cyanide.

In order to capture the sulfur, lime stone or dolomite may be fed to the gasifier

$$\mathrm{CaO} + \mathrm{H_2S} \rightarrow \mathrm{CaS} + \mathrm{H_2O}\,. \qquad (16.41)$$

16.11.6 Optimal Coal Gasifiers

Gasifiers convert carbonaceous feedstock into gaseous products at high temperature and elevated pressure in the presence of oxygen and steam. Partial oxidation of the feedstock provides the heat. At operating conditions, chemical reactions occur that produce synthesis gas or *syngas*, a mixture of predominantly CO and H_2.

16.11.7 Classification of Gasifiers

A wide variety of gasifier designs has been developed for different applications and types of fuel used. The important parameters used for selecting the type of gasifiers are temperature, pressure, reactant gases, and the method of contacting. The different types of gasifier used in combined cycle technology are:

1. The entrained-flow (downflow) gasifier
2. The E-GAS entrained flow (up flow) gasifier
3. The Shell entrained flow (up flow) gasifier
4. The fluidized-bed gasifier
5. The transport reactor gasifier
6. The Lurgi dry ash gasifier
7. The British Gas/Lurgi fixed-bed gasifier
8. The Prenflo entrained bed gasifier

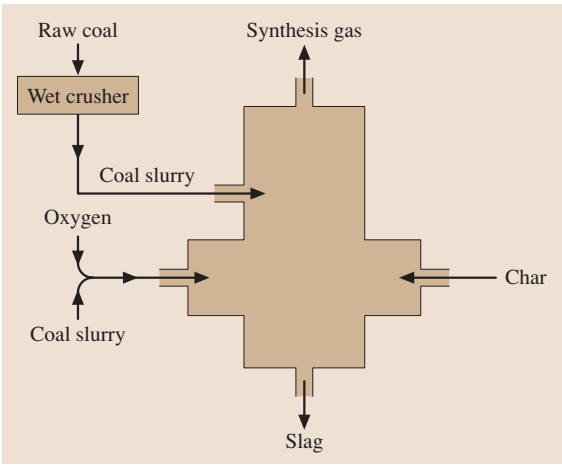


Fig. 16.15 E-GAS entrained-flow system◀

16.11.8 E-GAS Entrained Flow

Figure 16.15 shows the arrangement of the E-GAS entrained-flow gasifier. The gasification reaction occurs in two regions of the gasifier, normally called the primary and secondary. The raw feed coal is crushed in the wet crusher to produce slurries and is fed to the gasifier from the bottom portion normally called the first stage or the primary region of the gasifier. In this region exothermic gasification/oxidation reactions take place at a temperature of 1300–1400 °C, so great care has to be taken for the design. This region is normally lined with special slag-resistant refractory. There are two opposed horizontal burners with a horizontal cylinder provided in this region. Oxygen is used to gasify the slurry fed to the first stage of the gasifier. Usually 80% of the coal slurry is fed into this region. The remaining 20% of the coal slurry is injected into the hot raw gas coming out from the first stage.

The second stage contains a vertical cylinder that is perpendicular to the first stage. The endothermic gasification/devolatilization reactions in this stage reduce the final gas temperature and add some hydrocarbons to the product gas. Char is produced in the second stage. The hot gas leaving the gasifier is cooled in a fire-tube product gas cooler, generating saturated steam that is sent to the steam turbine.

16.12 Magnetohydrodynamic (MHD) Power Generation

16.12.1 Principle of MHD

The underlying principle of MHD power generation is elegantly simple. An electrically conducting fluid is driven by a primary energy source (e.g., the combustion of coal or a gas) through a magnetic field, resulting in the establishment of an electromotive force within the conductor in accordance with the principle established by Faraday [16.6].

Furthermore, if the conductor is an electrically conducting gas, it will expand, and hence the MHD system constitutes a heat engine concerning an expansion from high to low pressure in a manner similar to that of a gas turbine. Unlike gas turbine which involves surface interaction, the MHD system, however, involves a volume interaction between a gas and the magnetic field through which it is passing.

16.12.2 General Characteristics

The MHD generator can properly be viewed as an electromagnetic turbine because its output is obtained from the conducting gas–magnetic field interaction directly in electrical form rather than in mechanical form. Electrical conduction in gases occurs when electrons are available to be organized into an electric current in response to an applied or induced electric field. The electrons may be either injected or generated internally, and, because of the electrostatic forces involved, they require the presence of corresponding positive charge from ions to maintain electrical neutrality. An electrically conducting gas consists in general of electrons, ions to balance the electric charge, and neutral atoms or molecules; such a gas is termed a plasma [16.7].

16.12.3 The Production of Plasma

In MHD generators, electrons to support the flow of current can be obtained by two different methods: thermal and nonequilibrium ionization.

16.12.4 Thermal Ionization

Plasma is obtained by heating the gas to a sufficiently high temperature to yield electrons through ionization.

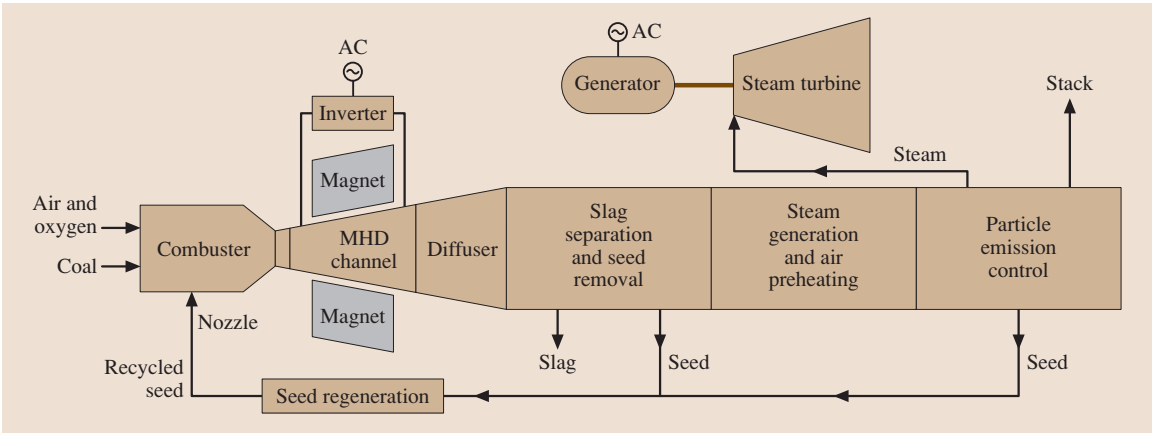


Fig. 16.16 Schematic view of an MHD steam power plant

16.12.5 Nonequilibrium Ionization

Sufficiently strong electric fields are induced in a manner similar to that in gas discharge devices. In either case, the mechanism of energy transfer from the flowing fluid to the electrical output can be thought of as a coupling of the electron-containing gas to the ions through electromagnetic forces. The ions in turn are embedded in the background of an atomic or molecular gas and lack mobility by virtue of their being coupled to the molecules or ions through collision processes described by kinetic behavior.

16.12.6 MHD Steam Power Plant

A schematic diagram of an MHD steam power plant is shown in Fig. 16.16. The product gas with a high electrical conductivity formed by burning the coal in the combustor is mixed with potassium carbonate (called seed) to increase its conductivity. The ionized gas then flows through a strong magnetic field, inducing an electric field and setting up a potential difference between the walls of the duct. Then using a solid-state inverter, the direct current (DC) generated is converted to alternating current (AC).

After flowing through the magnetic field, the hot gasses are then used to generate steam and turn a turbine as in a conventional plant. As the heat is transferred, slag is removed for disposal and the seed is captured for recycling. One considerable drawback is the need to use expensive superconducting magnets that must be cooled to −269 °C (4 K) to generate the necessary magnetic fields. About 50% of efficiency can be achieved if the MHD generator is operated in tandem with a conventional plant.

16.13 Total-Energy Systems for Heat and Power Generation

16.13.1 Cogeneration

A plant producing both electrical power and process heat simultaneously is called a cogeneration plant. The generation process can be any amalgamation of two different forms of useful energy (electricity and heat). Loss of energy in the condenser is very well utilized to obtain the required amount of heat, as the efficiency of a cogeneration system is higher than a conventional system operating in condensing mode. The heat obtained from a cogeneration plant is used for space heating of buildings, drying, to produce hot water or steam, or in various industrial processes.

The different types of cogeneration technologies are:

1. Steam turbine (ST) cogeneration system
2. Gas turbine (GT) cogeneration system
3. Combined cycle gas turbine (CCGT) cogeneration system
4. Internal combustion engine cogeneration system

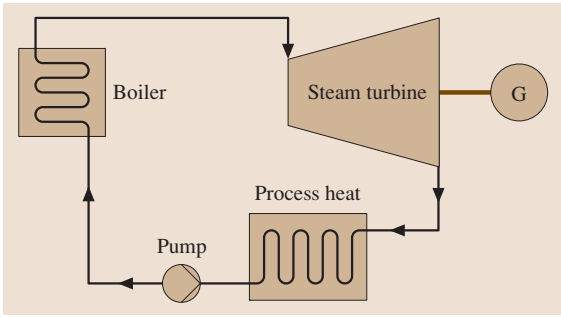


Fig. 16.17 Layout of a back-pressure turbine

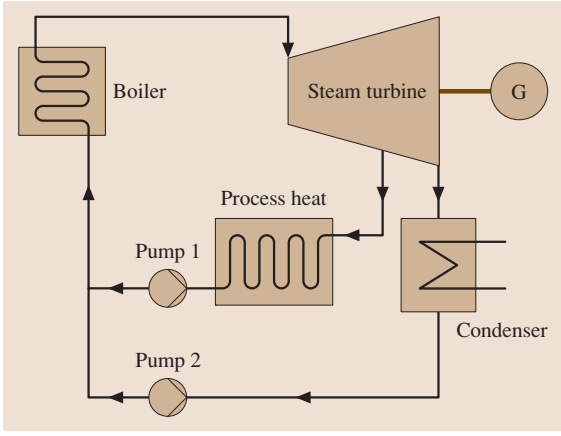


Fig. 16.18 Extraction condensing turbine

Steam Turbine Cogeneration System

Figure 16.17 depicts the layout of a cogeneration with a back pressure turbine. This signifies generation of electricity and heat by means of steam, generated in boilers by burning a suitable fuel (fossil or nonfossil, e.g., biomass). The steam is sent to the back-pressure extraction turbine, where it is expanded until its designed back pressure. The steam turbine coupled with electric generators drives the steam turbine and produces electricity. Normally in cogeneration the required amount of heat is obtained by energy from the outlet of the steam turbine, from the back-pressure outlet or from the extractions of the steam turbine depicted in Figs. 16.17 and 16.18. The required thermal energy is delivered in the form of steam at a pressure corresponding to the design or with the required temperature in accord with the thermal energy level.

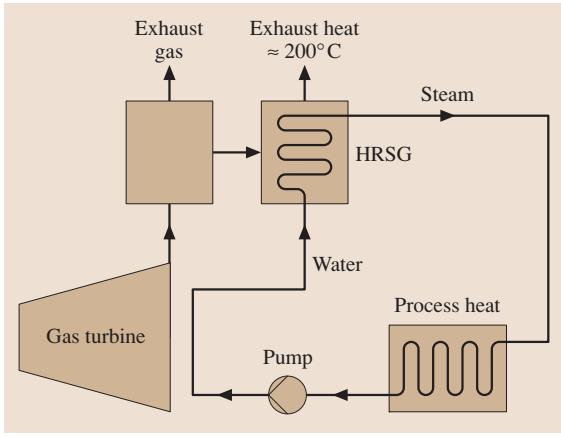


Fig. 16.20 Layout of an internal combustion engine cogeneration system

Gas Turbine Cogeneration System

In this design, a combination of gas and steam turbines is installed to convert energy from fuel to mechanical energy to drive electric generators.

Figure 16.19 represents the layout of a gas turbine cogeneration system. Hot gases from the gas turbine are

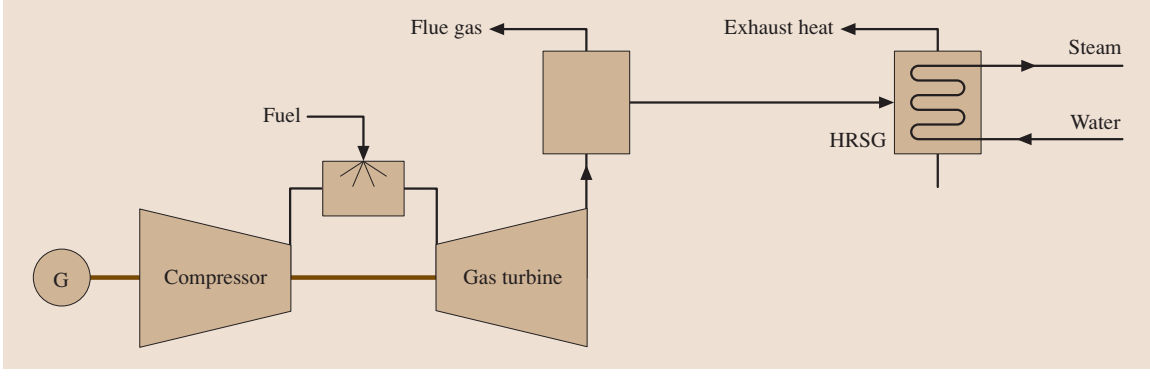


Fig. 16.19 Layout of a gas turbine cogeneration system

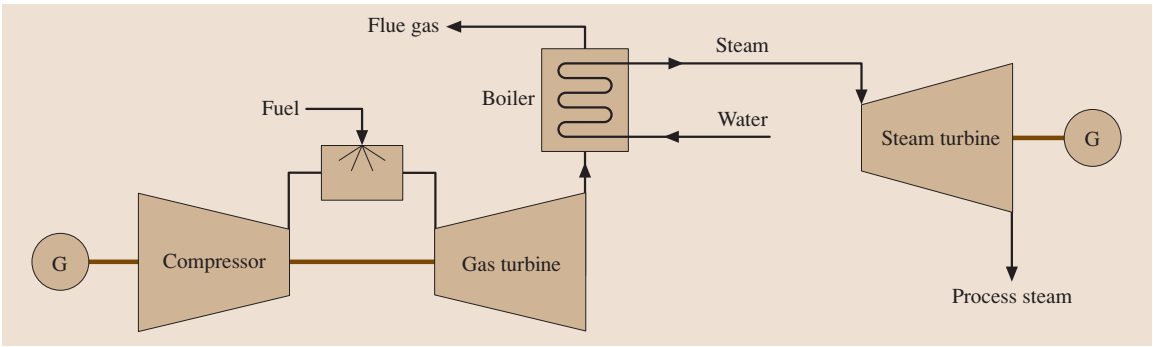


Fig. 16.21 Layout of a combined-cycle cogeneration system

used to generate steam for the steam turbine. Heat energy of the steam escaping from the turbine is used to provide heat. Mechanical energy (gas turbine and steam turbine) is converted into electrical energy with the help of a generator. The net efficiency of the plant is more than 50%.

Internal Combustion Engine Cogeneration System

Combined heat and power (CHP) systems are also available to serve smaller-sized facilities. In this type of facility, these smaller *modular* cogeneration units can generate 20–650 kW, and produce hot water from engine waste heat. Figure 16.20 represents the layout of an internal combustion engine cogeneration system. The generator converts mechanical work produced at the engine shaft into electrical energy. Heat resulting from combustion during power generation is used for process heat supply or heating purposes. Exhaust gases and the engine cooling water function as heat sources.

Combined Cycle Cogeneration System

Figure 16.21 represents the layout of a combined cycle cogeneration system. The combined cycle cogeneration system consists of a gas-turbine-driven generator package, a heat recovery steam generator closely matched to the process steam conditions, and a steam turbine matched to the HRSG's output and connected to the generator.

The advantages of this design are:

1. High thermal efficiency
2. Operating flexibility
3. Low installation cost

16.14 Transformation of Regenerative Energies

16.14.1 Wind Energy Power Plant

Wind is essentially created by solar heating of the atmosphere. Wind as a power source is attractive because it is plentiful, inexhaustible, and nonpolluting. Furthermore, it does not impose an extra heat burden on the environment [16.8]. Unfortunately, it is nonsteady and unreliable. Control equipment has been devised to start the wind power plant whenever the wind speed reaches 30 km/h. Methods have also been found to generate constant-frequency power with varying wind speeds and consequently varying speeds of windmill propellers. Wind power may prove practical for small power needs in isolated sites, but for maximum flexibility, it should be used in conjunction with other methods of power generation to ensure continuity [16.9]:

1. Small generators (0.5–10 kW) for isolated single premises
2. Medium generators (10–100 kW) for communities
3. Large generators (1.5 MW) for connection to the grid

Figure 16.22 depicts the arrangement of a wind turbine. The wind power is a measure of the energy available in the wind and is a function of the cube (the third power) of the wind speed. If the wind speed is doubled, the power in the wind increases by a factor of eight, so small differences in wind speed lead to large differences in electric power. This example points out that minor differences in wind speed due to either site selection or measurement errors can have a major bearing on a decision to invest in a wind turbine. For this

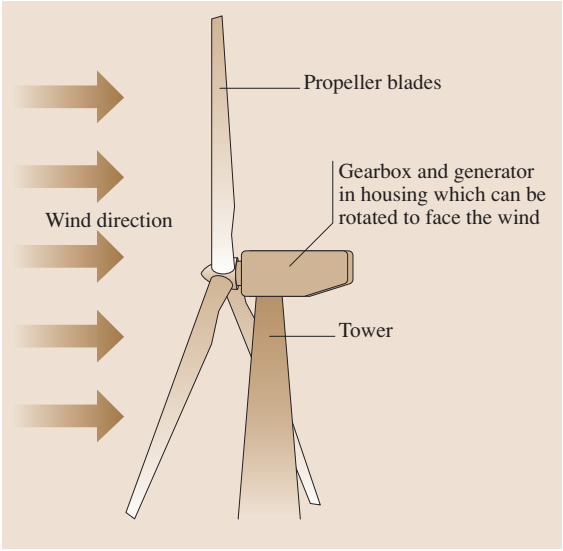


Fig. 16.22 Schematic view of the conversion of wind energy into electrical power▶

reason, a thorough and accurate wind study is imperative before buying a wind turbine.

The theoretical power in a wind stream is given by

$$P = 0.5\rho A V^3 \quad \text{(W)}\,, \tag{16.42}$$

where ρ is the density of air (1201 g/m^3 at normal temperature and pressure (NTP), V is the mean air velocity (m/s), and A is the swept area (m^2).

This equation states that the power is equal to one-half of the air density times the rotor area, times the cube of the wind speed. Moreover the air density varies due to the following features:

1. Elevation
2. Temperature
3. Weather fronts

16.15 Solar Power Stations

The sun acts as an atomic furnace as it converts mass into a huge amount of energy, according to Einstein's mass–energy relation $E = mc^2$. Every second it converts over 657 million tons of hydrogen into 653 million tons of helium, producing energy by nuclear fusion. The remaining 4 million tons of mass is discharged into space as energy. The Earth accepts only about one two-billionths of this. The radiation energy from the sun is massive, within 15 min the sun radiates energy equivalent to the amount of energy mankind consumes during a whole year. It is imperative to utilize this energy radiated by the sun, in order to fulfill our energy requirement.

16.15.1 Significant Features of Solar Energy

Ultimately solar energy is free and results in no hazard to the environment. In sunny countries, solar power can be used where there is no easy way to supply electricity to a remote place. It is also convenient for low-power devices such as solar-powered garden lights and battery chargers. There are various ways to produce electricity from the energy obtained from the sun.

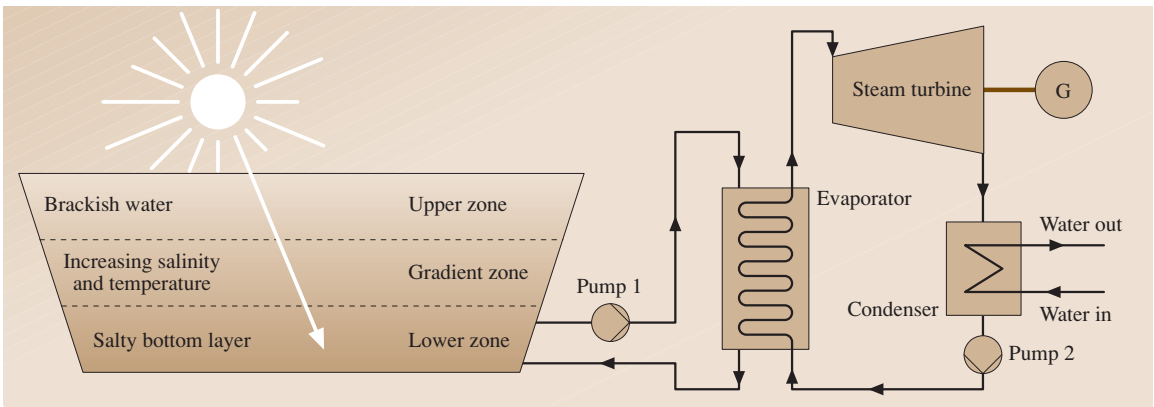


Fig. 16.23 Schematic of a solar pond

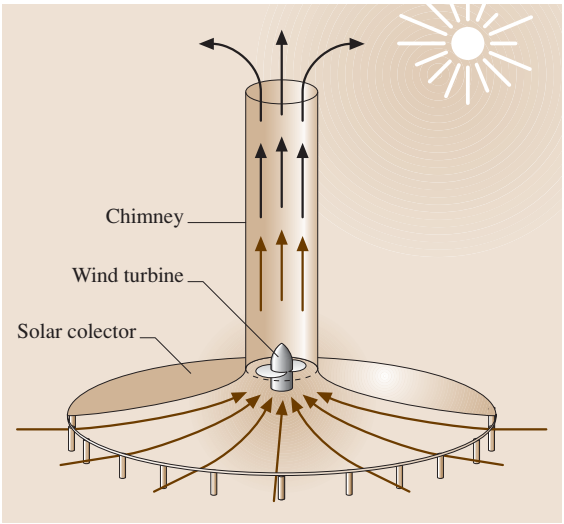


Fig. 16.24 Arrangement of a solar chimney

16.15.2 Solar Cells or Photovoltaic Cells

A solar panel is made up of the natural element silicon, which becomes charged electrically when subjected to sunlight. Sunlight is composed of packets of energy called photons. These photons contain various amounts of energy corresponding to the different wavelengths of light. The solar cell is made up of a material with well-defined possible energy levels for electrons. When a photon is absorbed in the active material, the energy of the photon is transferred to an electron, which becomes excited from a low to a higher energy level. The asymmetric structure of the solar cell ensures that this electron escapes from its normal position, carrying away the extra energy. This electron leaves the cell through a metal contact and becomes part of the current in an electrical circuit.

16.15.3 Solar Pond

A solar pond is a relatively low-tech, low-cost approach to harvest solar energy. A solar pond consists of three layers of water with different salt concentrations, as shown in the Fig. 16.23. A low-salt-content salt layer and an intermediate layer with a salt gradient, which creates a density gradient that averts heat exchange by natural convection in the water. The bottom layer with a higher salt content reaches a temperature of around 90 °C. The heat energy from the salty bottom layer can be used to generate electricity.

Normally an organic fluid with lower boiling point is used to convert this low-grade energy into electricity. The hot brine is pumped from the salty bottom layer and is sent into the evaporator, where it transfers heat to the organic fluid and the properly utilized brine is again sent to the salty bottom layer. The organic fluid in vapor form at the exit of the evaporator is used to rotate the turbine and generator.

The exhaust from the turbine is condensed in the condenser and then pumped back to the evaporator and hence the cycle is closed. The efficiency of the overall system entirely depends on the salinity and purity of the pond, and normally it is quite difficult to maintain salinity and dirt-free condition.

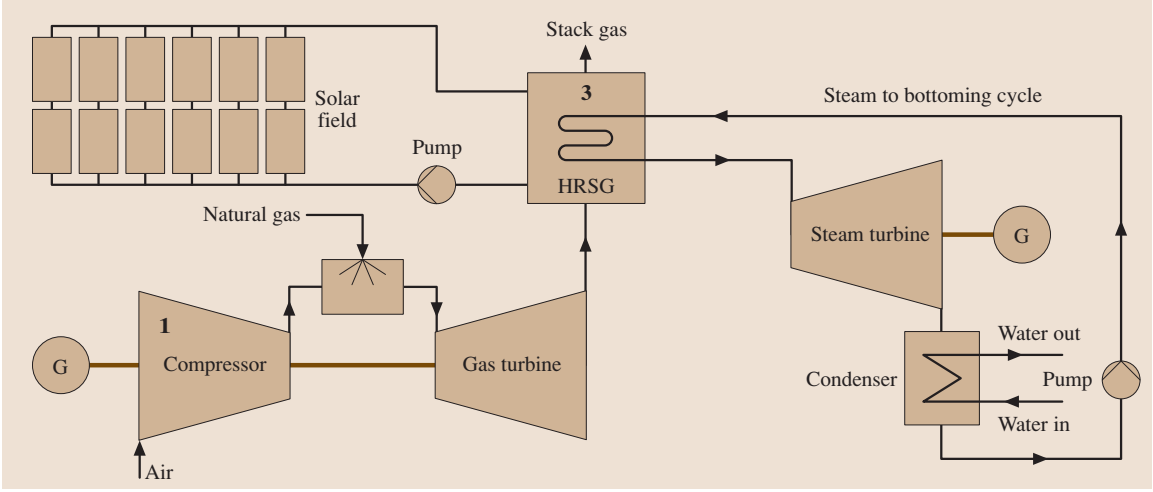


Fig. 16.25 Layout of an integrated solar combined cycle using a solar field

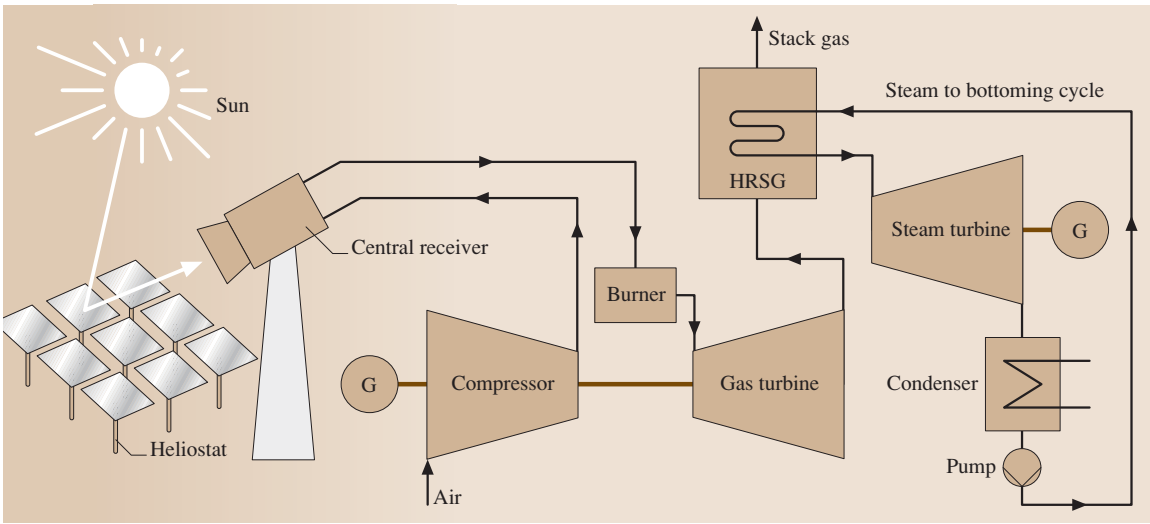


Fig. 16.26 Layout of integrated solar combined system using heliostats

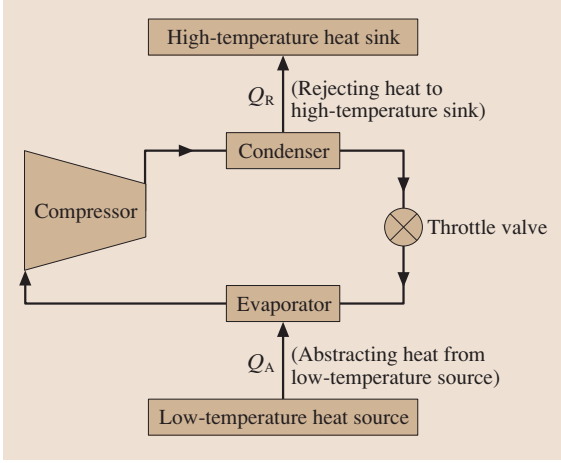


Fig. 16.27 Layout of a heat pump

16.15.4 Solar Chimney

Solar chimney technology is an innovative cost-effective technique to harness energy from the sun. A solar chimney consists of a glass collector, a wind turbine, an electrical generator, and a chimney. The glass collector is situated at the bottom and extends to a height of 2–7 m above the ground, as shown in the Fig. 16.24. The chimney is located at the center and its height varies from 500 to 1000 m. The conversion efficiency depends on the chimney height, the collector area, and the intensity of solar radiation.

The working principle of a solar chimney is quite simple: the air trapped inside the glass collector is heated by the radiation from the sun and becomes less dense, ascending the chimney. The rising hot air turns the wind turbine, which generates electricity.

16.15.5 Integrated Solar Combined Cycle Power System

Solar energy from a parabolic trough integrated in a combined cycle leads to high efficiency and low emission. Figure 16.25 shows the layout of this integrated solar combined cycle system. The heat from the parabolic trough collector can be directly utilized in the heat-recovery steam generator in addition to the heat supplied from the gas turbine exhaust. This integration seeks to achieve efficient operation even though solar energy intensity varies according to weather and time of day.

Sun-tracking concentrating mirrors called *heliostats* collect the sunlight and concentrate it onto a central receiver located at the top of a high tower, as shown in the Fig. 16.26. The receiver converts the radiative solar energy to thermal energy by heating air to the temperature required for direct feeding of the gas turbine.

Other possibilities for the exploitation of this energy can be endothermic chemical processes, or direct use of thermal energy for process heating. Central receiver technology is a promising alternative technology [16.10] overcoming the disadvantages of the trough technology.

16.16 Heat Pump

A reversed Carnot cycle can be used as heat pump. If the aim is to heat a body or space, the heat is rejected at a high temperature to the body or space and the heat is absorbed at a lower temperature from the ambient air or circulating water. Thus heat is drawn from the atmospheric air and pumped to the space to be heated. Such a cycle is called a heat pump cycle (Fig. 16.27) and the coefficient of performance (COP) for a Carnot reversed cycle heat pump is given by

$$\mathrm{COP}_{\text{Carnot heat pump}} = \frac{\text{Heat rejected}}{\text{Work done}} ,$$

$$\mathrm{COP}_{\text{Carnot heat pump}} = \frac{T_2}{T_2 - T_1} ,$$

where T_1 and T_2 are temperatures of source and sink.

The efficiency of this device to transfer heat Q_R to a high-temperature body is

$$\text{Efficiency} = \frac{\text{Energy effect sought}}{\text{Energy input}} .$$

16.17 Energy Storage and Distribution

Energy storage plays an important role in the competent management of energy resources. The demand for electricity fluctuates with time, which affects the economics of power plants that are normally designed for higher capacity.

The ultimate aim of an energy storage device is to reduce the economic losses due to fluctuating demand. When the demand is lower than the capacity, energy is stored. When the demand is higher than the capacity, the stored energy is released. This will provide savings in operating cost and ensure complete customer satisfaction, which can improve the status of the organization in the international market. Finally, energy storage is commonly used in stand-alone applications, where it can serve as an uninterruptible power supply (UPS) unit. The most important energy storage technologies are:

1. Pumped hydro power
2. Compressed energy storage
3. Flywheels
4. Electrochemical storage devices
5. Thermal energy storage devices
6. Secondary battery energy storage

16.17.1 Pumped Hydro Power

Pumped hydro facilities consist of two large reservoirs, one located at the base level and the other located at a different elevation. In pumped hydro, surplus power is utilized to pump the water from the lower reservoir to the upper reservoir, where it can be stored as potential energy. During periods of higher demand, water is sent back into the lower reservoir, passing through hydraulic turbines that generate electrical power [16.11]. The only drawback in pumped hydro power devices is that their construction cost is very high.

The combined efficiency of a pumped hydro system is given by

$$\eta_{\text{comb. effic.}} = \frac{\text{Total energy output}}{\text{Total energy input during a charge–discharge cycle}} . \tag{16.43}$$

16.17.2 Compressed Air Energy Storage

Excess energy is used to compress air and store it an airtight underground storage cavern. The stored energy is then released during periods of peak demand by expansion of the air through an air turbine. Three types of reservoirs can be used to store compressed air: salt caverns, aquifers, and hard rock caverns. When air is compressed for storage, its temperature increases according to

$$T_2 = T_1 \left(\frac{p_2}{p_1}\right)^{\frac{n-1}{n}} , \tag{16.44}$$

where n is the polytropic index, and P_1, T_1 and P_2, T_2 are the pressures and temperatures before and after compression.

Various studies have concluded that compressed air energy storage is competitive with combustion turbines and combined-cycle units, even without taking into account its unique benefits in terms of energy storage [16.12].

The layout of compressed energy storage device is shown in the Fig. 16.28. The heat of compres-

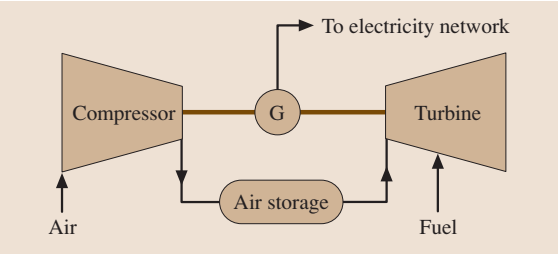


Fig. 16.28 Layout of compressed air storage

sion may be retained in the compressed air. This is called adiabatic storage and results in high storage efficiency, since more energy is recovered by expansion.

16.17.3 Energy Storage by Flywheels

Flywheel energy storage is a promising technology for providing intermediate energy storage. A flywheel storage device consists of a flywheel that rotates at a very high velocity and an integrated electrical apparatus that can operate either as a motor to turn the flywheel and store energy or as a generator to produce electrical power on demand using the energy stored in the flywheel. The use of magnetic bearings and a vacuum chamber helps reduce energy losses.

In a flywheel, energy storage is in the form of mechanical kinetic energy. The rotating mass stores the energy input so that the rotation can be maintained at a fairly constant rate. There are two main sources of losses in the flywheel: windage and bearing.

The energy stored in flywheel is

$$\mathrm{KE}_{\mathrm{disc}} = \frac{1}{4} M r^2 \omega^2 \,, \tag{16.45}$$

where M is the mass of the flywheel, and r is the radius of the flywheel.

In order to achieve high energy density, the rotation speed ω must be very high.

16.17.4 Electrochemical Energy Storage

Electrochemical energy storage is one of the recent technologies, which can be classified into three categories: primary batteries, secondary batteries, and fuel cells. These devices convert stored chemical energy into electrical energy. Primary and secondary batteries utilize the chemicals built into them, whereas fuel cells use chemically bonded energy supplied from the outside in the form of synthetic fuel [16.13].

16.17.5 Thermal Energy Storage

Thermal energy storage is ideally suited for applications such as space heating, where a low quantity of heat is required. The two distinct thermal energy storage mechanisms are sensible heat storage and latent heat storage. In sensible heat storage, energy can be stored as sensible heat by virtue of a rise in temperature of the storage medium.

16.17.6 Secondary Batteries

Large-scale battery use is almost unfeasible and their use is limited to battery-powered vehicles and storage for local fluctuating energy sources such as windmills or solar. The most widely used battery is the lead-acid battery, invented by Plante in 1859. The sodium-sulphur battery (200 W h/kg) and other combinations of materials are also being developed to obtain more output and storage per unit weight [16.9].

16.18 Furnaces

The furnace is the heart of a steam generation system. It is an enclosed chamber in which heat is produced by burning fuel, to heat water in the case of a steam generation system. Its dimensions and geometry are adapted to the amount of heat release, type of fuel, and to the method of firing so as to promote complete burning of the combustible and suitable disposal of the resulting ash. A furnace can provide combustion of fuel in solid, liquid or gaseous form.

Based upon the type of fuel used furnace are classified into:

1. Solid-fuel furnaces
2. Liquid-fuel furnaces
3. Gas-fuel furnaces

16.18.1 Combustion

Combustion or burning is a chemical process, an exothermic reaction between a substance (the fuel) and

a gas (the oxidizer), usually O_2, to release heat. The presence of CO_2 in the product gas signifies complete combustion whereas CO signifies incomplete combustion.

The basic chemical equations for complete combustion are

$$C + O_2 \rightarrow CO_2 \,, \tag{16.46}$$

$$2H_2 + O_2 \rightarrow 2H_2O \,. \tag{16.47}$$

When the amount of oxygen supplied is insufficient for complete combustion then carbon will be burned incompletely with the formation of carbon monoxide

$$2C + O_2 \rightarrow 2CO \,. \tag{16.48}$$

The most important parameter to estimate the effectiveness of combustion is called the combustion efficiency, which depends on the following parameters:

1. Air–fuel ratio
2. Fuel–air mixing
3. Flame temperature
4. Flame shape
5. Fuel residence time
6. Degree of atomization (for liquid fuel)
7. Degree of turbulence

The theoretical air–fuel ratio for complete combustion is known as the stoichiometric ratio. In practice every oxygen molecule does not come into contact with a fuel molecule. In order to ensure complete combustion some amount of air is used to compensate this shortage of oxygen molecules, normally called *excess air*, and attain complete combustion. Turbulence enhances the proper mixing of fuel and oxygen and hence the combustion efficiency.

16.18.2 Ideal Combustion

The generalized ideal combustion equation can be written as

$$C_X H_Y S_Z O_K + \left(X + \frac{Y}{4} + Z + \frac{K}{2}\right) O_2 + mH_2O \rightarrow PCO_2 + QH_2O + GSO_2 \,. \tag{16.49}$$

The generalized combustion equation can be written

$$C_X H_Y S_Z O_K + \left(X + \frac{Y}{4} + Z + \frac{K}{2}\right) O_2 + 3.76\left(X + \frac{Y}{4} + Z + \frac{K}{2}\right) N_2 + mH_2O \rightarrow PCO_2 + QH_2O + RN_2 + GSO_2 \,. \tag{16.50}$$

The air–fuel ratio can be written

$$\frac{A}{F} = \frac{m_{\text{air}}}{m_{\text{f}}} \,. \tag{16.51}$$

16.18.3 Theoretical Dry Air–Fuel Ratio

1 kmole of oxygen and 3.76 mole of nitrogen generate a mixture of 4.76 mole of air.

The molecular weight of air is therefore

$$\frac{32 + 3.76 \times 28}{4.76} = 28.84 \,, \tag{16.52}$$

$$\left(\frac{A}{F}\right)_{\text{TD}} = \frac{m_{\text{air}}}{m_{\text{f}}} = \frac{4.76\left(X + \frac{Y}{4} + Z - \frac{K}{2}\right) 28.84}{100} \,. \tag{16.53}$$

16.18.4 Theoretical Wet-Air–Fuel Ratio

Ambient air is always humid in nature so the calculation of the theoretical wet-air–fuel ratio depends on the relative humidity. The specific humidity of air, w (kg) of moisture per kg of dry air.

Relative humidity

$$\phi = \frac{p_{\text{vapor,act}}}{p_{\text{vapor,sat}}} \,. \tag{16.54}$$

Amount of moisture in ambient air

$$n = \frac{4.76\,(X + Y/4 + Z - K/2)\,28.84w}{18} \,. \tag{16.55}$$

Theoretical wet-air–fuel ratio

$$\left(\frac{A}{F}\right)_{\text{TW}} = \frac{m_{\text{wet,air}}}{m_{\text{f}}} = \frac{\left[4.76\left(X + \frac{Y}{4} + Z - \frac{K}{2}\right)\right](28.84 + w18)}{100} \,. \tag{16.56}$$

16.18.5 Pressure Conditions

Not all boiler furnaces are airtight, especially stokers (boilers that burn solid fuels). Flue gases may escape into the plant area if the furnace pressure is greater than the atmospheric pressure. Other furnace designs may require draft and furnace pressure control. Typically, the furnace pressure is controlled using a balance draft system. The induced draft fan is modulated to maintain the

furnace at a slight negative pressure. Furnace pressure measurement and induced fan control is required. The location of the furnace pressure transmitter is important because the pressure is not uniform within the furnace. If the combustion airflow rate measurement is available, furnace pressure control can be more effective.

16.18.6 Emission

The combustion of fuel finally leads to the emission of various gases and particulate matter. The amount and chemical components of these emissions depend on the fuel type, boiler type and size, and the firing method. Different forms of emissions are described below.

16.18.7 Particulate Emissions

The particulates present in the stack gases depend primarily on the combustion efficiency and on the amount of ash contained in the fuel. All fuels except natural gas contain some quantity of ash or noncombustible material, which forms the majority of these particulates [16.14, 15].

16.18.8 Nitrogen Oxide Emission

The level of nitrogen oxides (NO_x) present in the stack gas depends on many variables; the furnace heat rate levels, temperature, and excess air are the major variables that affect NO_x emission levels. NO_x is one of the contributors to acid rain and ozone formation, visibility degradation, and human health concerns. Combustion of any fossil fuel generates some level of NO_x due to the high temperature and availability of oxygen and nitrogen from both the air and fuel. Based on the method of formation, NO_x can be classified as thermal NO_x and fuel NO_x.

16.18.9 Thermal NO_x

High-temperature oxidation (above 1200 °C) initiates the formation of NO_x, normally called thermal NO_x. The nitrogen and oxygen in the air dissociate at higher combustion temperatures and lead to the formation of NO_x. Thermal NO_x formation is typically controlled by reducing the peak and average flame temperature. Apart from a higher temperature, the formation of NO_x is also due to longer residence time and oxygen concentration. Three possible reactions for the formation of NO_x during combustion are [16.15]

$$N_2 + O \rightarrow NO + N\,, \tag{16.57}$$

$$N + O_2 \rightarrow NO + O\,, \tag{16.58}$$

$$N + OH \rightarrow NO + H\,. \tag{16.59}$$

16.18.10 Fuel NO_x

Fuel NO_x refers to the formation of chemically bound nitrogen in the fuel during combustion. The fuel–air ratio is one of the deciding factors for the formation of fuel NO_x. Conversion of fuel-bound nitrogen to NO_x is strongly dependent on the fuel–air ratio but is relatively independent of the combustion-zone temperature. The formation of NO_x happens at two levels, one is during oxidation of volatile nitrogen and another is from the char during combustion.

16.18.11 Sulfur Dioxide Emission

SO_2 is an acidic gas formed by the combustion of sulfur in the fuel with oxygen. Dilute sulfuric acid is a major constituent of acid rain. An aqueous solution of sulfurous acid (SO_3) is formed when sulphur dioxide combines with water. This can easily oxidize in the atmosphere to form sulfuric acid (H_2SO_4).

16.18.12 Solid-Fuel Furnaces

Furnaces that use solid fuels for combustion are normally called solid fuel furnaces. Fuels include, coal, coke, and firewood (wood chips and pellets).

16.18.13 Stokers and Grates

There are various ways to introduce coal into the furnace. Stokers play a vital role in distributing coal into the furnace. Stokers are normally differentiated on the basis of how the coal is introduced into the fire. Different types of stokers are:

1. Traveling-grate stoker
2. Chain-grate stoker
3. Spreader stoker
4. Vibrating stoker
5. Underfeed stoker

Traveling-grate stokers have been in use for the past 50 years and are the most popular way to burn coal in stokers for boilers. In addition to coal, traveling-grate

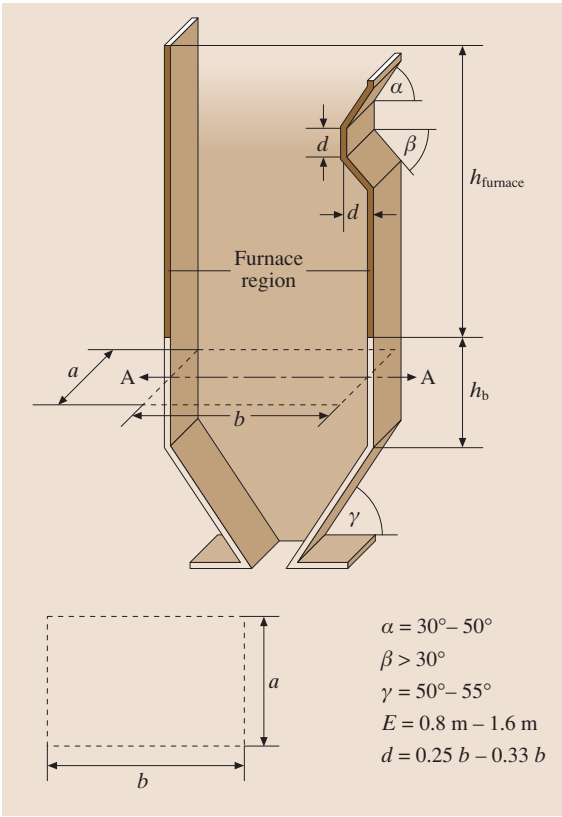


Fig. 16.29 Basic furnace geometry

stokers can handle a wide variety of waste fuels. Underfeed stokers are only used for small plants, and mainly for steam production rather than power generation. An underfeed stoker pushes the coal up into the furnace from below the grate.

Chain and traveling grate furnaces have similar characteristics. Coal lumps are fed continuously on to a moving grate or chain. Air is drawn through the grate and through the bed of coal on top. As the coal enters, it is heated by radiation from the refractory. Moisture and volatile matter are driven off. The chain/grate moves the coal slowly into the region in which ignition is established, and the temperature in the coal bed rises. The carbon gradually burns off, leaving ash which drops off at the end into a receptacle, from which it is removed for disposal. The ash formed may have a carbon content as high as 4–5 %.

In the spreader stoker arrangement, a high-speed rotor throws the coal into the furnace over a moving grate to promote fuel distribution.

16.18.14 Pulverized-Fuel Furnaces

The furnace is the most important part of a boiler. Its primary function is to provide sufficient space for fuel particles to burn completely and to cool the flue gas to a temperature at which the convective heating surfaces can be operated safely. Structurally the boiler furnace consists of the combustion space surrounded by water walls. The principle of combustion in pulverized-fuel boilers is that coal ground into a fine powder is mixed with air and transported into a combustion chamber where it is burnt in a flame similar to the flame of a liquid or gaseous fuel.

The furnace is designed to perform two functions simultaneously, namely:

- The release of the chemical energy of fuel by combustion: the first task of combustion technology is to burn the fuel efficiently and steadily, to consume controlled excess air (as little as possible).
- To generate a flame with a controlled shape that will generate the lowest amount of pollutants and ensure the transfer of heat from the furnace to the working fluid inside the water walls. The important task of furnace heat removal is to produce a controlled furnace exit gas temperature (FEGT). The FEGT is an important aspect of boiler safety.

The combustion gases leave the furnace at a safe temperature which will not cause clinkering to the subsequent heating surface.

A furnace can be characterized geometrically by its linear dimensions: the front width a, the depth b, and height h_f as mentioned in Fig. 16.29 which are estimated according to the rated fuel consumption and the thermal, physical, and chemical properties of the fuel to be used.

The furnace height should be sufficiently high that the flame should not heat the superheater tubes. Furnace width and depth are two of the most important parameters for design. The minimum value of furnace depth depends on the capacity of the boiler and types of fuel burnt.

The following factors influence the width and depth of the furnace:

1. The arrangement of burners
2. The heat release rate per unit furnace area
3. The power output of each burner
4. The flame length

Depending upon the condition of the ash leaving the bottom of the furnace, pulverized furnaces can be classified into two types:

1. Dry-bottom furnaces
2. Wet-bottom furnaces

16.18.15 Dry-Bottom Furnace

Dry bottom means that the boiler has a furnace bottom temperature below the ash melting point. In this furnace design ash or slag is removed in the dry state. In the dry-bottom type, ash falling down from the boiler furnace is conveyed by a continuously moving scraper chain conveyor to the clinker grinder, which is conveyed to the slurry sump through sloping trenches with the help of a high-pressure water jet. The furnace hopper dimension is one of the most important parameters for the effective removal of dry ash from the furnace. Normally the walls of the hopper are inclined at an angle of 48–60°.

16.18.16 Wet-Bottom Furnace

In a wet-bottom furnace ash or slag is removed in the wet state. In this design the hopper is filled with water and ash falling down is quenched and removed after a predetermined time with the help of a jet pumping system and conveyed to the slurry sump. This type of hopper is known as a water-impounded bottom ash hopper. There are two types of wet-bottom boilers: the slag-tap boiler and the cyclone boiler. The slag-tap boiler burns pulverized coal and the cyclone boiler burns crushed coal. In each type, the bottom ash is kept in a molten state and tapped off as a liquid. Both boiler types have a solid base with an orifice that can be opened to permit the molten ash that has collected at the base to flow into the ash hopper below. The ash hopper in wet-bottom furnaces contains quenching water. When the molten slag comes into contact with the quenching water, it fractures instantly, crystallizes, and forms pellets.

16.19 Fluidized-Bed Combustion System

Fluidized-bed combustion (FBC) technology is a very appropriate technology for efficient solid-fuel combustion. The velocity at which the bed behaves like a fluid is called fluidization velocity. At this point the pressure drop across the bed is equal to the weight of the particles per unit cross section of the bed. Increasing the air velocity imparts turbulent motion, which helps ensure proper mixing of the gas and particles.

Solid coal is crushed in the crusher to the required size, normally a mean particle diameter of around 0.5–12 mm, and fed into the fluidized bed, where it mixes with gas and inert materials and undergoes various reactions such as drying, devolatilization, combustion of volatiles, and the combustion of residual tar. During drying all the moisture present in the coal will be removed and the volatile matter is gradually released as the temperature increases; moreover the release of volatile matter is directly proportion to the temperature and inversely proportional to the gas pressure. After the volatile is released, combustion of the volatiles takes place, and finally burning of char takes place depending on the coal type, fluidization system, and char diameter.

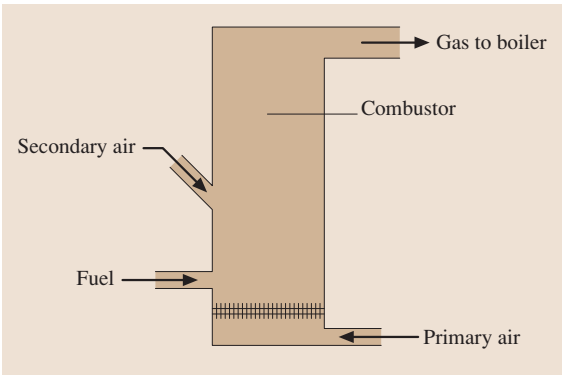


Fig. 16.30 Bubbling fluidized-bed combustion

Fluidization technique results in a vast improvement in combustion efficiency of high-moisture-content fuels, and is adaptable to a variety of waste-type fuels. The scrubbing action of the bed material on the fuel particle enhances the combustion process by shredding away the carbon dioxide and char layers that normally form around the fuel particle. FBC reduces the amount of sulfur emitted in the form of SO_x emissions. Limestone is used to precipitate out sulfate during combustion. FBC boilers can burn fuels other than coal, and the lower temperatures of combustion (800 °C) have other added benefits as well. FBC

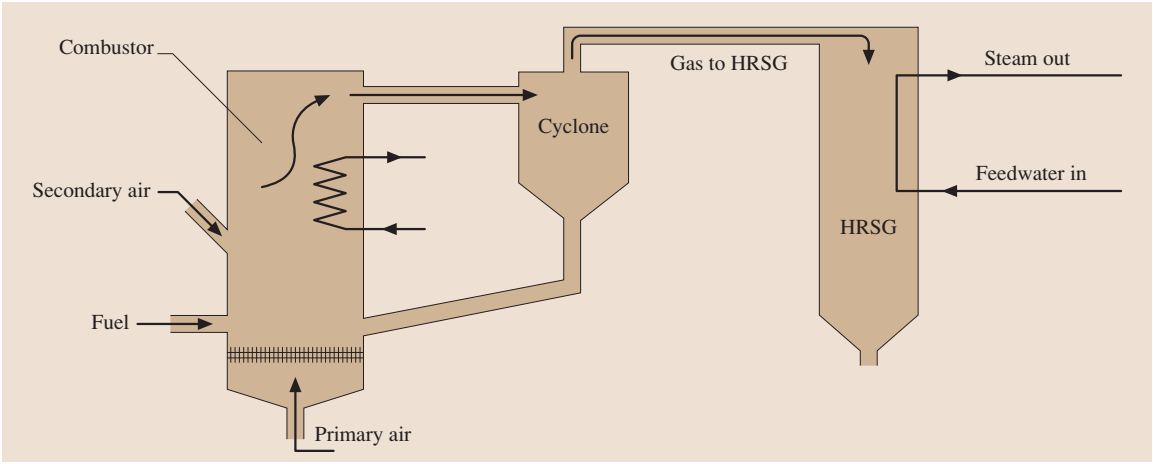


Fig. 16.31 Circulating fluidized-bed combustion

in boilers at atmospheric pressure can be particularly useful for high-ash coals, and/or those with variable characteristics.

Features of Fluidized-Bed Combustion Systems

Emissions from fluidized-bed combustor are always lower than conventional combustion technologies for the following reasons:

- Low combustion temperatures and low excess air within the fluidized bed reduces the formation of NO_x.
- High combustion efficiency results in flue gases that contain low amounts of CO.
- Emission such as SO_x may be abated within the fluidized-bed system by injecting limestone into the bed and ammonia into the vapor space.
- Fluidized-bed combustion is an environmentally favorable, proven technology for the disposal of solid wastes and the generation of energy.

16.19.1 Bubbling Fluidized-Bed Combustion

Bubbling beds use a low fluidizing velocity, so that the particles are held mainly in a bed with a definable surface. Inert materials are often used to improve the bed stability, together with limestone for SO_2 absorption. In-bed tubes are used to control the bed temperature and generate steam.

16.19.2 Circulating Fluidized-Bed Combustion

Generally circulating fluidized-bed combustion uses a boiler and a high-temperature cyclone. The gas velocity is as high as 4–8 m/s. Coarse fluidizing medium and char in the flue gas are collected by the high-temperature cyclone, and are recycled to the boiler as shown in Fig. 16.31. Air is introduced into the bed in two regions. About 40–70% of the air is injected through the nozzle grate at the bottom of the bed, normally called the primary air, and the remaining air, called the secondary air, is injected through nozzles on the side walls of the fluidized bed. The combustion proceeds in two zones: a primary reducing zone in the lower section of the combustor and an oxidizing zone in the upper part of the combustor where complete combustion is achieved through the use of excess air. This staged combustion, at controlled low temperatures, effectively suppresses NO_x formation. To increase the thermal efficiency, a preheater for fluidizing air and combustion air, and a boiler feedwater heater are installed.

16.20 Liquid-Fuel Furnace

Liquid fuels such as gasoline, kerosene, and diesel fuel are used directly for combustion in a liquid-fuel-fired furnace. Two major categories of fuel oil are burned by combustion sources: distillate oil and residual oil. Distillate oils are commonly used in domestic and small commercial applications, and include kerosene and diesel oil. Residual oil is mainly used in utility and large commercial application.

16.20.1 Special Characteristics

1. Due to the finer atomization (mean particle size 20 mm in the cause of fuel oil) better combustion takes place, resulting in a reduction in excess air requirement from 31% to 6%.
2. Liquid fuels having higher calorific value when compared to solid fuels.
3. These fuels occupy less space during storage.
4. These furnaces have practically no ash formation.

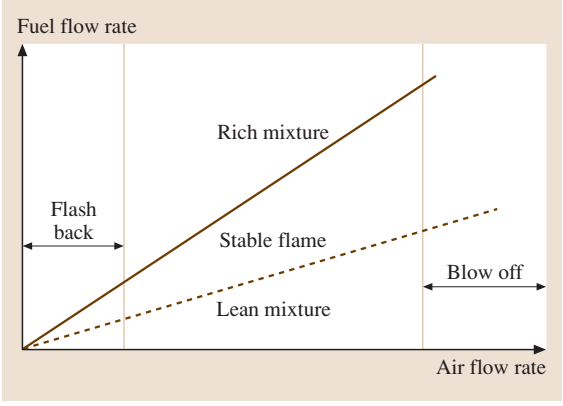


Fig. 16.32 Stability and flammability limits

5. The furnaces are easy to light up and shut down.
6. The furnaces have higher efficiency.
7. The fuel consumption rate can be controlled easily.

16.21 Burners

In order to achieve efficient combustion proper mixing of air and fuel is always necessary. The fuel must be evenly dispersed in the combustion airstream such that the fuel and air can make intimate contact. Failure to achieve this results in unburnt or partially burnt fuel. It is very important to design a component to achieve this task and provide better combustion efficiency. The burner is the apparatus for burning fuels continuously and more securely. The burner design attempts to achieve this using a variety of techniques. Important design criteria of burners are the burning rate, burning velocity, flashback, and the quenching diameter. Among these, the burning velocity is the deciding factor for the performance of the burner. It is defined as the relative velocity of the flame front to the unburned gases which is propagating normal to the flame front. These design parameters are related as follows

$$\text{burning velocity} > \text{flow velocity: flashback limit},$$
$$\text{burning velocity} < \text{flow velocity: blow-off limit},$$
$$\text{burning velocity} = \text{flow velocity: stable flame}.$$

The graph between the fuel flow rate and the airflow rate describes the mixing conditions, as shown in Fig. 16.32. The amount of air inflow affects the heat of the flame, and can be controlled by adjusting the slot openings. Stoichiometric calculation of the designed fuel determines the amount of air required, the fuel gas produced, and the heat released per unit weight of fuel burnt.

Minimum Air Required per Kilogram of Solid or Liquid Fuel Burnt

If a known mass of fuel consists of fractions of carbon, hydrogen, oxygen, and sulphur, then the minimum air required to burn the fuel is calculated as

For carbon requirement

$$= \text{mass of carbon (kg)} \times \left(\frac{8}{3\,\text{kg}}\right) \text{ of oxygen.}$$

For hydrogen

$$= \text{mass of hydrogen (kg)} \times 8\,\text{kg of oxygen.}$$

For sulphur

$$= \text{mass of sulphur} \times 1\,\text{kg of oxygen}.$$

Therefore, the amount of oxygen required for complete combustion of 1 kg of fuel is given by

$$\tfrac{8}{3}\text{C} + 8\text{H}_2 + \text{S} - \text{O}_2$$
$$= \tfrac{8}{3}\text{C} + 8\left(\text{H}_2 - \frac{\text{O}_2}{8}\right) + \text{S}$$

and atmospheric air contains 23 wt % of oxygen. Therefore, the minimum amount of air required per kilogram of fuel is

$$\frac{100}{23}\left(\frac{8}{3}\mathrm{C}+8\mathrm{H_2}+\mathrm{S}-\mathrm{O_2}\right)$$
$$=11.6\mathrm{C}+34.8\left(\mathrm{H_2}-\frac{\mathrm{O_2}}{8}\right)+4.35\mathrm{S}\,.$$

This is the theoretical amount of air required. The quantity of air in excess of this theoretical minimum that is actually required for complete combustion of the solid and liquid fuels is called the excess air.

16.21.1 Various Types of Burners

Burners can be classified in a number of ways: principle of operation (swirl type, parallel, or direct flow type), fuel (gaseous-fuel burners, liquid-fuel burners, and solid-fuel burners), and geometry (orifice burners, nozzle burners, flat flame burners, and burners with different shapes of openings).

16.21.2 Liquid-Fuel Burners

Fuel oil is the commonly used liquid fuel in the burners. An oil supply pump is provided to supply oil from the tank to the burner. To facilitate the fuel vaporization oil burners are designed to increase the contact surface area of the oil with air. To facilitate this, oil is atomized before entering the combustion chamber. It is difficult to atomize the fuel at ambient temperature so it has to be heated. this is achieved by single unit heating loops and a centralized heating system. A liquid fuel burner consists of a fuel supply line with a fuel atomizer, a concentric primary air supply with an air bag, and a secondary air supply with a vane control valve.

An oil burner is a mechanical device that combines fuel oil with appropriate amounts of air before delivering the mixture to the point of ignition in a combustion chamber. It is essential for the efficiency of the combustion process that the oil–air mixture is well homogenized and with as few pure droplets of fuel oil as possible. Fuel oil burners either vaporize or atomize the fuel oil.

Fuel oil burners can in general categorized into:

- Gun-type (atomizing) burners (pressure gun)
- Pot-type (vaporizing) burners
- Rotary-type fuel oil burners

16.21.3 Gun-Type Burners (Pressure Gun)

A gun-type burner atomizes the fuel oil by forcing the oil through a nozzle and spraying it into a gun-like airflow atomic nozzle. The liquid forms microscopic particles or globules that are well mixed and partly evaporated before being ignited in the combustion chamber. A residential gun-type burner normally requires a 551–896 kPa oil pressure. Commercial and industrial burners require 689–2068 kPa. The gun type is very flexible and can be used within a large range of

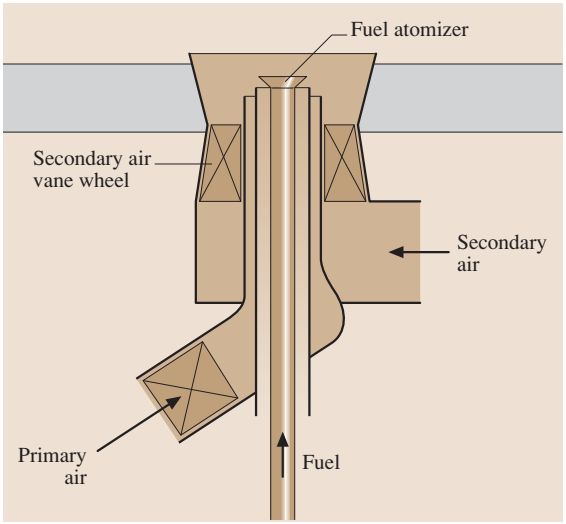


Fig. 16.33 Liquid-fuel burner arrangement

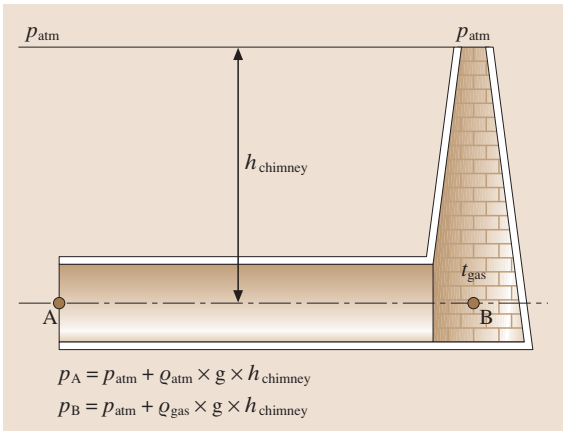


Fig. 16.34 Natural draught produced by a chimney

applications, from relative small residential heaters to larger industrial heating applications.

16.21.4 Pot-Type Burners

In a pot-type fuel burner the fuel evaporates into the combustion air. They are in general:

- Natural draft burners
- Forced draft burners
- Sleeve burners

In atmospheric pot-type heaters gravity causes the oil to flow to the burner. The natural draft burner relies on the natural draft in the chimney for air supply. The forced draft burner relies on a mechanical fan and/or the chimney for air supply. The perforated sleeve burner is only used in small applications. The pot-type burner is the most inexpensive of the fuel oil burners and has the lowest operating cost. A disadvantage of the pot-type is its limited capacity. This type is in general most suitable for smaller applications.

16.22 General Furnace Accessories

16.22.1 Fans

A fan moves a quantity of air or gas by adding sufficient energy to the stream to initiate motion and overcome resistance to flow. It helps to supply air for combustion, drying coal, and recirculation of gas.

16.22.2 Forced Draft Fan

The forced draft fan supplies combustion air and is important because it ensures adequate air–fuel mixing and keeps the flame away from the nozzle. The furnace operates under pressure and flue gases are exhausted by forced draft convection. The airflow rate is typically controlled with air dampers. Some applications use fan speed control.

16.22.3 Induced Draft Fan

The induced draft fan exhausts flue gases from the furnace and induces combustion air into the furnace by having the furnace operate under negative pressure. An induced draft fan can handle higher-temperature gas, which may contain corrosive ash.

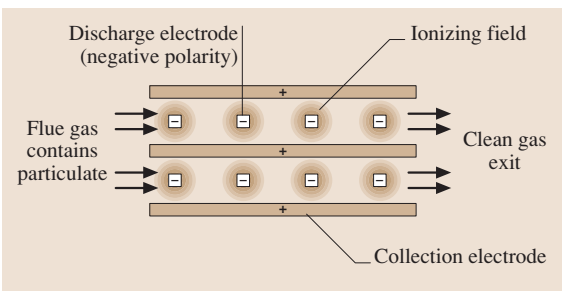


Fig. 16.35 Electrostatic precipitator (ESP), schematic diagram (top view)

16.22.4 Balanced Draft (BD)

The balanced draft design uses both forced draft and induced draft fans. It is used when the furnace design requires draft and furnace pressure control. Balance draft furnaces are typically operated at a slightly negative pressure, but they can also operate at slightly positive pressure. A furnace pressure measurement and induced fan control are required.

16.22.5 Primary Air Fans

In a coal-fired boiler, primary fans are used to supply the air needed to dry the coal and transport it to the boiler. Primary fans are usually located before or after the preheater.

16.22.6 Stacks

Stacks are used to create the required pressure differential for the flow of air and flue gas.

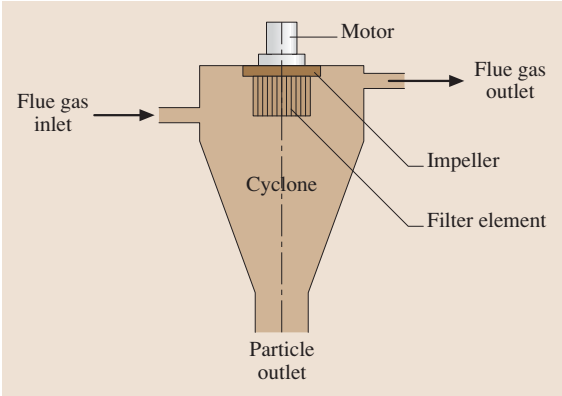


Fig. 16.36 Arrangement of a cyclone

The draft is the difference between the atmospheric pressure and the static pressure of the combustion gases in a furnace gas passage. In a coal-fired power plant tow methods are used to create a draught in the furnace:

1. A natural draft, or
2. A mechanical draft

16.22.7 Natural Draft

Natural draft establishes furnace breathing by continuous exhalation of flue gas and continuous inhalation of fresh air (Fig. 16.37). The tall chimney creates the natural draught by the temperature difference between the hot gases in the chimney and the cold atmospheric air outside the chimney.

The advantages of this method are:

- No external power is required.
- Air pollution is less since gases are discharged at a high level.
- No maintenance cost.
- Capital cost is less than artificial draught.

The draught or pressure difference produced is given by

$$\Delta p_{\mathrm{nd}} = \pm(\rho_{\mathrm{a}} - \rho_{\mathrm{g}})gH\,, \tag{16.60}$$

where Δp_{nd} is the head of natural draft (Pa), ρ_{a} is the ambient air density (kg/m^3), ρ_{g} is the gas density in the flue (kg/m^3), and H is the height difference between the beginning and the end of the section (m).

The flue gas density ρ_{g} is calculated as

$$\rho_{\mathrm{g}} = \rho_{\mathrm{g}}^{0}\frac{273}{273+T_{\mathrm{g}}}\,, \tag{16.61}$$

where T_{g} is the gas temperature (°C), ρ_{g} is the gas density in the flue under standard atmospheric conditions (1 atm) kg/m^3

$$\rho_{\mathrm{g}}^{0} = \frac{1-0.01A+1.302\alpha V^{0}}{V_{\mathrm{g}}}\,, \tag{16.62}$$

where α^0 is the excess air ratio in the flue gas, V^0 is the theoretical air requirement for unit weight of fuel (N m^3/kg), V_{g} is flue gas produced per unit weight of fuel (N m^3/kg $\approx \alpha V^0$), and A is the percentage ash content of the fuel.

The stack provides two functions:

1. Assisting the fan in overcoming pressure losses
2. Helping disperse the gas effluent

The amount of flow is limited by the strength of the draft. The pressure variation inside the chimney differs from atmospheric pressure. The variation of chimney pressure depends on the temperature variation along the chimney, which itself depends on the rate of cooling of hot gas due to natural convection.

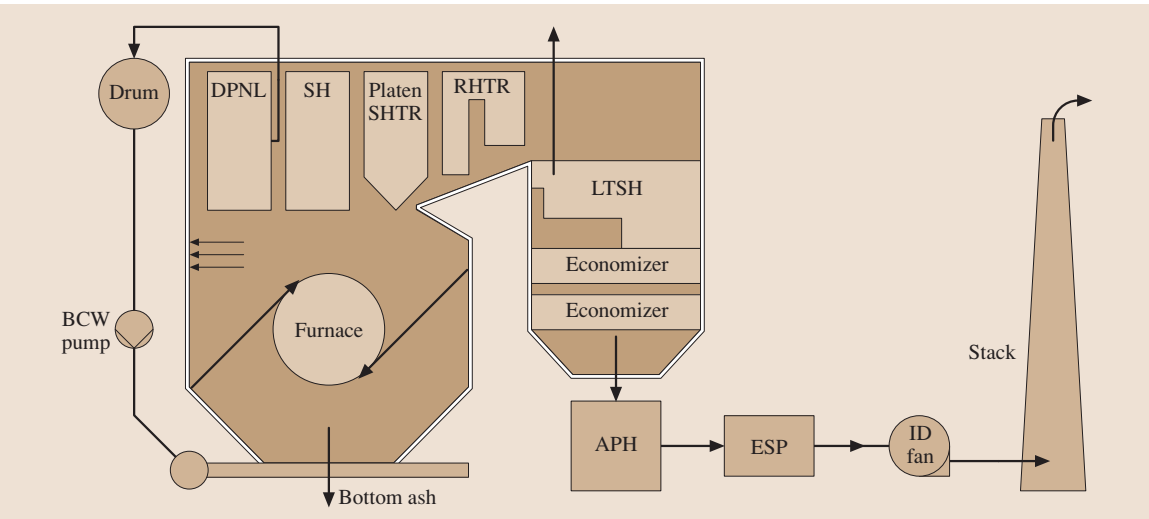


Fig. 16.37 Thermal structure of a boiler furnace (APH – air preheater, EPS – electrostatic precipitator, ID fan – induced draft fan, LTSH – latent superheater, platen SHTR – platen superheater, RHTR – reheater,SH – superheater)

16.22.8 Artificial Draught

In modern power plants, the draught should be flexible to meet the needs of fluctuating loads and it be independent of atmospheric conditions. To achieve this, the aid of draft fans becomes mandatory; by employing the fans, the height of the chimney can also be reduced. There are two types of fans used for producing mechanical draught:

1. Forced draught (FD)
2. Induced draught (ID)

16.22.9 Forced Draught

In this system, the blower (forced draft fan) is located at the base of the boiler near the grate. Air is forced into the furnace by the forced fan and the flue gases are forced to chimney through the economizer and air preheater.

Advantages of the Forced Draught System

- Since the fan handles cold air, the fan size and the power required are lower.
- No need for water-cooled bearings because the air being handled is cold.
- The pressure throughout the system is above atmospheric pressure so that air leakage into the furnace is reduced.

16.22.10 Induced Draught

In an induced draught system, a blower (induced draft fan) is placed near (or) at the base of the chimney. The fan sucks the flue gas from the furnace, creating a partial vacuum inside the furnace. Thus atmospheric air is induced to flow through the furnace to aid the combustion of fuel. The flue gases drawn by the fan passes through the chimney to the atmosphere.

16.22.11 Balanced Draught

In the induced draught system, when the furnace is opened for firing, cold air enters the furnace and dilates the combustion. In the forced draught system, when the furnace is opened for firing, the high-pressure air will try to blow out suddenly and the furnace may stop. Hence the furnace cannot be opened for firing and inspection in both systems. Balanced draught, which is a combination of induced and forced draught, is used to overcome these difficulties.

16.23 Environmental Control Technology

There are many technologies that can be used in industry to reduce the emissions of pollutants to the atmosphere, and these can be applied before, during, or after combustion.

16.23.1 Particulate Emission Control

There are several types of equipment available to control particulate matter from the flue gas which includes:

1. Electrostatic precipitators
2. Fabric filters
3. Mechanical collectors
4. Venturi scrubbers

16.23.2 Electrostatic Precipitators

When the ash particles present in the flue gas pass through the electrostatic precipitators (ESP) at a certain velocity, they become charged electrically and are attracted towards the collecting plate, which is normally positively charged.

Figure 16.35 shows a schematic diagram of ESP. The particulate-laden gas, normally laden with flyash, is sent through pipes with negatively charged plates which give the particles a negative charge. The particles are then routed past positively charged plates, or grounded plates, which attract the newly negatively charged ash particles. The particles stick to the positive plates until they are collected. The air that leaves the plates is then clean of harmful pollutants. Velocity is one of the important factors that affect the performance of an electrostatic precipitator. A lower velocity allows more time to collect the ash particles.

16.23.3 Fabric Filters

Fabric filters are used to remove particles from the gas stream. They are made up of woven or felted material. Fabric filters are generally in the form of a cylindrical bag. Fabric filters generally operate in a temperature range of 120–180 °C. The choice between ESP and fabric filtration generally depends on coal type, plant size, and boiler type and configuration. The two fundamen-

tal parameters in sizing and operating bag houses are the air-to-cloth (A/C) ratio (m/s) and the pressure drop (mm water gauge, Pascal or in H_2O). In operation, dust-laden gas flows through the filters, which remove the dust particles from the gas stream.

The most important factors that affect the performance of fabric filters are:

1. Flue gas temperature
2. Dew point and moisture content
3. Particle size distribution
4. Chemical composition of the fly ash

Fabric filters are classified into three types:

1. Pulse jet fabric filters
2. Reverse-air fabric filters
3. Shake-deflate filters

16.23.4 Pulse Jet Fabric Filters

Pulse jet fabric filters use high-pressure air to clean the filter bags, and are provided in standard configurations that are capable of treating gas flow rates up to about 300 000 ACFM (actual cubic feet per minute). Custom-designed units can handle larger flow rates.

16.23.5 Shake-Deflate Filters

This kind of filters collect the dust inside the bags as in the reverse-air design. To clean the bags, the top ends are shaken by a driver linkage.

16.23.6 Reverse-Air Fabric Filter

The reverse-air fabric filter is a customized design for utility boilers and industrial applications where large volumes of process gas flow (250 000 ACFM and more) must be cleaned. The systems consist of 6–24 structural compartments. Compartments are available with nominal 20 or 30 cm diameter bags with typical bag lengths of 7.31–11 m.

16.23.7 Mechanical Collectors

Mechanical dust collectors are often called cyclones. Cyclones are used to remove dust and fibrous material either as the first stage of a scrubber or fabric filter system. Although cyclones are an established form of dust collector, care and application knowledge are required to ensure correct sizing. The arrangement of a cyclone separator is shown in Fig. 16.36. The basic principle is the centrifugal force created by spinning a gas stream, which is used to separate the particles from the gas. In a conventional cyclone, the gas enters a cylinder tangentially, where it spins in a vortex as it proceeds down the cylinder. A cone section causes the vortex diameter to decrease until the gas reverses on itself and spins up the center to the outlet pipe or vortex finder. A cone causes flow reversal to occur sooner and makes the cyclone more compact. Dust particles are centrifuged toward the wall and collected by inertial impingement.

The collected dust flows down in the gas boundary layer to the cone apex where it is discharged through an airlock or into a dust hopper serving one or more parallel cyclone.

16.23.8 NO_x Control

It is very important to control the level of NO_x emitted from power plants. NO_2 from the exhaust reacts with sunlight and hydrocarbons to produce photochemical smog and acid rain constituents. The following techniques are used to reduce the level of NO_x formation in current practices:

- Low excess air operation
- Off-stoichiometric combustion, combustion modification
- Flue gas recirculation and treatment

Low Excess Air Operation

This technique involves a reduction in the total quantity of air used in the combustion process. By using less oxygen, the amount of NO_x produced is reduced.

Off-Stoichiometric Combustion

This technique involves the mixing of the fuel and air in a way that reduces the peak gas temperatures and peak oxygen concentrations. Advanced low-nitrogen-oxide burners can reduce emissions by up to 30%. Such burners can be installed in either new or existing combustion plants. For a low-NO_x burner, sudden heating up and temperature rise is important. In this case the high-temperature zone is very close to the burner compared with a conventional burner, so the pulverized coal is heated very rapidly in order to increase the fractional volatile and nitrogen release according to quantity introduced. Also, the recirculation flow near the center of the burner is important, because hot gas returning to the burner creates a very high-temperature region at this point. Altogether, the modified shape of the flow divider

and the pulverized fuel nozzle, together with an optimized strength of swirl in the air flow, produce a strong internal recirculation and NO_x reducing zone in the CI-a burner, with longer residence time in this region and reduced unburnt matter. So the result is a reduction of 50% in unburnt carbon and, at the same time, a significant reduction of at least 10% in NO_x production.

Over Fire Air (OFA)

This technique keeps the mixture fuel rich and completes the combustion process using air injection nozzles.

Combustion Modifications

Flue Gas Recirculation. A second method is to add some of the flue gas with the combustion air at the burner, normally known as flue gas recirculation. This increases the gas weight which must be heated by the chemical energy in the fuel, thereby reducing the flame temperature.

Flue Gas Treatment

Selective Noncatalytic Reduction Systems (SNCR). This technique involves the injection of ammonia (NH_3) or urea into the hot gas zone where reactions leading to reduction of nitrogen oxides can occur. The reactions are completed within the boiler, and no waste products are generated. There is a risk of ammonia (NH_3) being emitted into the atmosphere if temperatures are too low, however. SCNR systems are capable of reducing nitrogen oxides by 20–60%. The reactions are

$$4NH_3 + 4NO + O_2 \rightarrow 4N_2 + 6H_2O\,,$$
$$4NH_3 + 5O_2 \rightarrow 4NO + 6H_2O\,.$$

Flue Gas Desulphurization.

Precombustion Sulphur Control Technology. Removing the sulphur before burning is one of the challenging options. There are a verity of techniques available to reduce the sulphur, including coal scrubbing and oil desulphurization.

Another removal process is to change the design of the boiler and to install pressurized fluidized-bed combustors (FBC), which remove sulphur from the coal during the burning process.

Another process that removes sulphur dioxide from coal during combustion is the integrated gasification combined cycle. Coal is gasified under pressure with a mixture of air and steam which results in the formation of gas which can then be burned to produce electricity.

Post-Combustion Sulphur Control Technology. One of the post-combustion sulphur control methods (removing sulphur after burning) is flue gas desulphurization (FGD). In FGD processes, waste gases are scrubbed with a chemical absorbent such as limestone to remove sulphur dioxide. There are many different FGD processes, the main ones being the limestone–gypsum process and the Wellman–Lord regenerative process. Limestone–gypsum FGD involves mixing limestone and water with the flue gases to produce slurry which absorbs the sulphur dioxide. The slurry is then oxidized to calcium sulphate (gypsum) which can then be used in the building trade. FGD technologies can be classified into six main categories: wet scrubbers, spray dry scrubbers, sorbent injection processes, dry scrubbers, regenerable processes, and combined SO_2/NO_x removal processes.

16.24 Steam Generators

The steam generator is one of the main components in modern coal fired-power plants. Its concept, design, and integration into the overall plant considerably influence costs, operating behavior, and availability of the power plant. The thermal structure of the boiler furnace is shown in Fig. 16.37. Within the steam generator, fuel and air are forced into the furnace by the burner, where burning produces heat; from there fuel gas travels throughout the boiler, the feedwater absorbs the heat, and eventually absorbs enough energy to change into vapor. Boiler makers have developed various designs to extract the most energy from fuel and to maximize its transfer to the water.

Water enters the boiler, preheated, at the top as shown in the Fig. 16.37. The hot water naturally circulates through the tubes down to the lower area where it is hot. The water heats up and flows back to the steam drum, where the steam collects. Not all of the water is turned to steam, so the process starts again. Water keeps on circulating until it becomes steam. Meanwhile, the control system measures the temperature of the steam drum, along with numerous other readings, to determine if it should keep the burner burning, or shut it down. Sensors also control the amount of water entering the boiler, known as the feedwater. A steam generator is normally equipped with

basic component like a furnace, economizer, reheater, superheater, evaporator, air preheater, and auxiliary devices.

16.24.1 Types of Steam Generators

The classification of boilers depends on various phenomena, such as furnace position, the type of fuels used, tube contents, circulation etc.

16.24.2 Boiler Safety

Boiler safety is one of the prime aspects while operating the boiler. Operating the pressure above the design pressure is extremely dangerous, so proper control of the pressure inside the steam generator is very important. Though boilers are usually equipped with a pressure-relief valve, if the boiler fails to contain the expansion pressure, the steam energy is released instantly. This combination of exploding metal and superheated steam can be extremely dangerous.

The concentration of solids in the boiler should be measured and the boiler blow-down at such intervals as necessary to maintain established limits. Blow-down valves are placed at the lowest point of the boiler for the purpose of blowing sediment or scale from the boiler. They should be maintained in good working order and have to be opened and closed carefully when used.

Boilers should always be brought online slowly and cold water should never be injected into a hot system as sudden changes in temperature can warp or rupture the boiler. Because many boilers are fired by natural gas, diesel or fuel oil, special precautions need to be taken. Boiler operators should ensure that the fuel system, including valves, lines, and tanks, is operating properly with no leaks. The low-water cutoff is the most important electrical/mechanical device on a boiler for maintaining a safe water level. If a low-water condition develops, it could very well result in an overheating and explosion of the boiler. The low-water cutoff should be tested at least weekly.

To prevent furnace explosions, it is imperative that boiler operators purge the boiler before ignition of the burner. Workers should check the fuel-to-air ratio, the condition of the draft, and the flame to make sure that it is not too high and not smoky. Ventilation systems should also be inspected and maintained to make sure that combustion gases do not build up in the boiler room.

16.24.3 Boiler Water Treatment

Efficient performance of the boiler depends upon the quality of the water. The treatment of the boiler feedwater is required to prevent excessive fouling of the heat transfer equipment and the erosion of turbine blades.

The common impurities present in the raw water are:

1. Dissolved solids – calcium, magnesium
2. Suspended solids – mineral matter
3. Dissolved gases – oxygen and carbon dioxide
4. Scum-forming substances – carbonate, chlorate, and sulphate

In the steam boiler industry, high-purity feedwater is required to ensure proper operation of steam generation systems. High-purity feedwater reduces the use of boiler chemicals due to less frequent blow-down requirements. This lower blow-down frequency also results in lower fuel costs.

The boiler system loses water through steam and water leaks. Additional water called *make-up water* is added to the boiler to replace these losses. The amount of make-up water and the level of naturally occurring impurities in this water will determine the type of water treatment required. Boiler heating systems that have very few leaks require a simple water treatment program.

All water contains dissolved minerals and these minerals, if allowed to reach high enough levels in the boiler water, will come out of solutions and form as a hard shell on the hot surfaces of the boiler. This hard shell is called *scale* and is often found on the outside of the fire tubes or the inside of water tubes. Scale insulates the heating surfaces, reducing the ability of the fire tubes to transfer heat from the hot combustion to the boiler water. High stack temperatures or ruptured tubes are common problems related to scale build up. Boiler water also contains dissolved gases such as oxygen or carbon dioxide. These gases, in the presence of water and metal, can cause corrosion. Corrosion eats away the metal, affecting the durability of the boiler.

For boiler feedwater treatment, depending on its requirements, a number of processes can be utilized including chemical treatment/lime softening, dual-media filtration, carbon adsorption, conventional reverse osmosis membranes, and final ion-exchange resin polishing. Various methods of pretreatment of water are discussed below.

Lime Soda Process

Calcium and magnesium salts are removed using lime and soda ash. The chemical reactions during this process are

$$\begin{aligned} Ca(HCO_3)_2 + Ca(OH)_2 &= 2CaCO_3 + 2H_2O\,, \\ MgCl_2 + Ca(OH)_2 &= Mg(OH)_2 + CaCl_2\,, \\ MgSO_4 + Ca(OH)_2 &= Mg(OH)_2 + CaSO_4\,, \\ CaSO_4 + Na_2CO_3 &= CaCO_3 + Na_2SO_4\,, \\ CaCl_2 + Na_2CO_3 &= CaCO_3 + 2NaCl\,. \end{aligned}$$

Insoluble components such as calcium carbonate and magnesium hydroxide settle at the bottom.

Deaeration

The feedwater from the condenser contains dissolved gases such as O_2 and CO_2. Oxygen is found in feedwater with a relatively high partial pressure, so, it requires a near saturation temperature to be disassociated from the water. Oxygen in combination with water will attack iron and cause corrosion. This reaction occurs in two steps

$$Fe^{2+} + 2OH^- = Fe(OH)_2\,.$$

Then,

$$\underset{\substack{\text{Soluble}\\\text{ferrous}\\\text{hydroxide}}}{4Fe(OH)_2} + \underset{\substack{\text{Dissolved}\\\text{oxygen}}}{O_2} + \underset{\text{Water}}{2H_2O} = \underset{\substack{\text{Insoluble}\\\text{ferric}\\\text{hydroxide}}}{4Fe(OH)_3}$$

A deaerator is an open-type feedwater heater, in which the steam that is bled from the turbine is mixed directly with water. When the water temperature increases, the dissolved gases reduce.

16.24.4 Shell-Type Steam Generator

This is a cylindrical boiler where the shell axis is vertical to the firing floor. Originally it consisted of a chamber at the lower end of the shell, which contained the combustion appliance. The gases rose vertically through a flue surrounded by water. Large-diameter (100 mm) cross tubes were fitted across this flue to help extract heat from the gases, which then proceeded to the chimney. Later versions had the vertical flue replaced by one or two banks of small-bore tubes running horizontally before the gases discharged to the chimney. The steam was contained in a hemispherical chamber forming the top of the shell.

The present vertical boiler is generally used for heat recovery from exhaust gases from power generation or marine applications. The gases pass through small-bore vertical tube banks. The same shell may also contain an independently fired section to produce steam at such times that there is insufficient or no exhaust gas available.

In relation to the thermal capacity generated, a shell-type boiler has much higher water contents than a water tube boiler. Therefore, a shell-type boiler is more robust towards load fluctuations or load demands that temporarily exceed the rated boiler capacity.

Shell-type boilers are fire-tube boilers, because the products of combustion pass through the boiler tubes. Lancashire and Cornish are examples of shell-type boilers.

16.24.5 Natural Circulation Boiler

The distinct features of natural circulation boiler are that natural circulation occurs due to the density difference between the fluids in the down comer and riser or is caused by convection currents that result from the uneven heating of the water contained in the boiler. The natural circulation has been largely used in boilers up to 140 bar. Based upon the position and geometry natural circulation boilers are classified into two types:

1. Vertical-tube type
2. Sloped-tube type

Figure 16.39 shows a typical water-tube natural circulation waste heat boiler with steam drum and down comer and riser pipes. Feedwater enters the drum from an economizer. This mixes with the steam/water mixture inside the drum. Down comers carry the resultant cool water to the bottom of the evaporator tubes while external risers carry the water–steam mixture to the steam drum. The heat transfer tubes also act as risers, generating steam.

The natural circulation is maintained due to the static head difference and natural convection due to the density differential between the mean down comer density and mean riser density. The down comers are located outside the furnace and away from the heat of combustion. They serve as pathways for the downward flow of relatively cool water.

The circulation ratio is defined as the ratio of the mass of steam–water mixture to steam generation. The natural circulation largely depends upon the

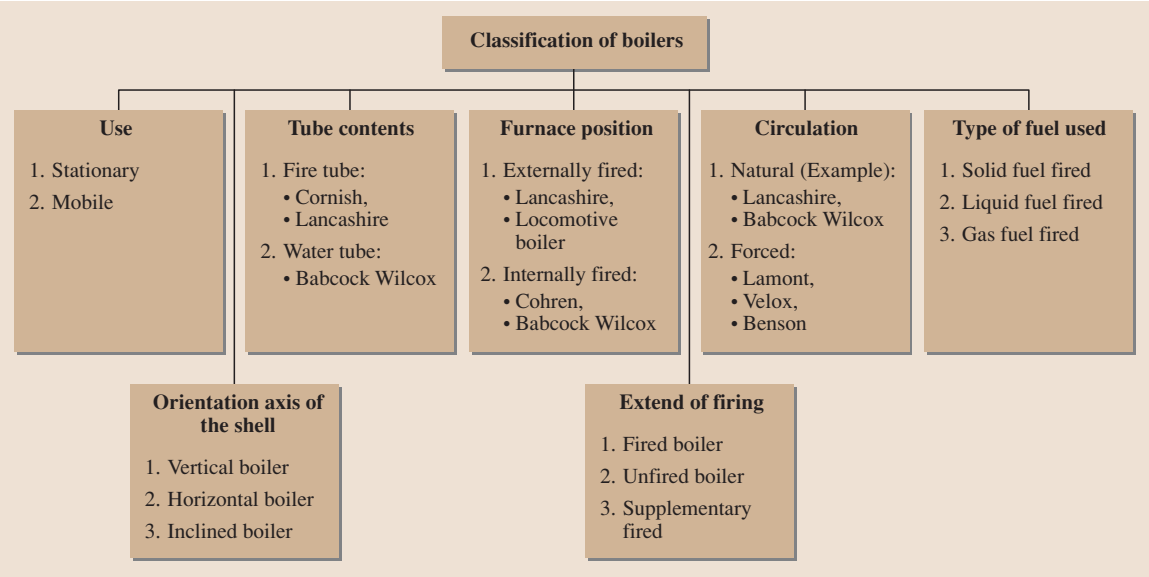


Fig. 16.38 Classification of boilers

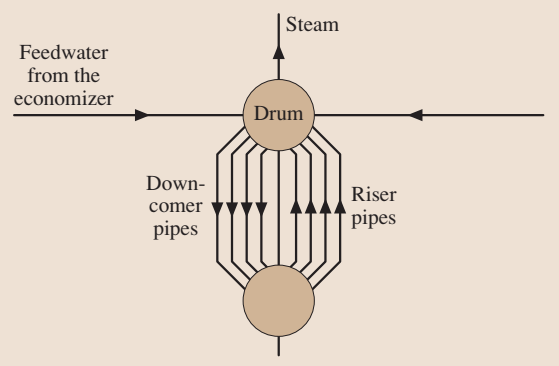


Fig. 16.39 Natural-circulation-type water tube boiler

height of the riser and the down comer, and the mean density of the fluids. The main feature of natural circulation boilers is the absence of water circulation pumps.

16.24.6 Forced Circulation Boiler

Modern boilers use the principle of forced circulation. When the boiler pressure is very high, the difference between the densities of water and the saturated steam decreases and hence less circulation occurs. The difference between the lower drum and steam drum must be increased in order to obtain the same output under higher-pressure condition. This can be achieved by means of forced circulation. The example of a forced circulation boiler is the Lamont boiler. The operation is basically similar to that of natural circulation unit, with a pump being used below the down comer for circulating water through the risers.

Normally forced circulation boilers are used if the pressure is very high. If the pressure is very high, the density difference between liquid/water and liquid mixture is very less, so, there abstract the problem of free flow like in natural circulation.

A once-through boiler is a forced circulation boiler. Once-through boilers have been favored in many countries for more than 30 years. They can be used up to a pressure of more than 30 MPa without any change in the process engineering. In the once-through boiler, subcooled water, supplied by a feedwater pump at the inlet of the tubes, is heated, evaporated, and superheated in the long evaporation tubes to produce superheated steam at a prescribed pressure and temperature at the outlet of the tubes.

A principle advantage of the once-through boiler is that it does not require circulating pumps or drums. The energy required for circulation is provided by the feed pump. Since the state of the superheated steam at the outlet is determined by the total amount of heat input, independent of the heat the heat flux distribution along the tube at a defined mass flow rate, the once-through boiler is a fairly simple device to obtain superheated steam of a prescribed state at any partial load.

16.24.7 Boiling Water Reactors

The fission zone is contained in a reactor pressure vessel, at a pressure of about 70 bar (7 MPa). At the temperature reached (approximately 290 °C), the water starts to boil and the resulting steam is produced directly in the reactor pressure vessel. After the separation of the steam and water in the upper part of the reactor pressure vessel, the steam is routed directly to a turbine driving an alternator. The steam coming out of the turbine is converted back into water by a condenser after having delivered a large amount of its energy to the turbine. It is then fed back into the primary cooling circuit where it absorbs new heat in the fission zone. Since the steam produced in the fission zone is slightly radioactive, mainly due to short-lived activation products, the turbine is housed in the same reinforced building as the reactor.

16.25 Parts and Components of Steam Generator

16.25.1 Superheaters

One of the most important accessories of a boiler is a superheater, which affects improvement and economy in the following ways. The steam that is produced in the boiler has a certain percentage moisture content. Due to the high velocities of the steam inside a turbine, the moisture content of the steam can erode the turbine blades. A superheater is utilized to remove the moisture content in the steam by raising the temperature while keeping the pressure constant. Steam that undergoes this process is referred to as superheated steam. Superheating improves the turbine internal efficiency and hence the lifetime of the turbine. The degree of superheating is a term which is used to describe the temperature difference between the raised temperature and the temperature at constant pressure.

A superheater therefore:

1. Increases the capacity of the plant
2. Reduces corrosion of the steam turbine
3. Reduces steam consumption of the steam turbine

Depending upon the way heat is transferred, superheaters are classified into three types:

1. Radiant superheaters
2. Convective superheaters
3. Combined radiative and convective superheaters

Convective superheaters are normally called primary superheaters and are located near the convective zone of the furnace, whereas radiant and combined superheaters are termed secondary superheaters.

Flow Arrangements of the Different Types of Superheater

The saturated steam from the drum is sent into the convective superheaters. After the convective superheater the steam is passed into the radiant superheater, where the heat is absorbed purely by means of radiation. Steam leaving the radiant superheater is sent into the desuperheater, where highly pure water is sprayed directly into the steam. The temperature of the steam leaving the pendent superheater should not exceed the rated value.

Superheaters are often divided into more than one stage such as:

1. A platen superheater
2. A pendent superheater
3. A horizontal superheater
4. A radiant superheaters

16.25.2 Radiant Superheater

Radiant superheaters receive energy primarily by thermal radiation from the furnace with little energy from convective heat transfer. Radiant superheaters are located at the furnace exit or turning section. The radiant superheater absorbs more enthalpy at partial loads when compared to the convective type. At lower loads the flow distribution inside the superheater tubes is less uniform. The radiant superheater outlet temperature decreases with increasing boiler output. At higher loads the mass flow rate of the combustion gas is high, because of increased amount of fuel and air for combustion. The convective heat transfer coefficient increases both inside and outside the tubes. Thus the steam receives more heat transfer per unit mass flow rate, and its temperature increases with load. The surface area required to transfer a given amount of energy will be lower due to the higher log mean temperature difference and higher heat transfer coefficient. Hence their cost may be lower in spite of the better grade of materials required.

16.25.3 Convective Heat Transfer

A convective superheater is located in a low-gas-temperature region ranging from 423 to 813 K lower, depending on the degree of superheating required. Since it is shielded by several rows of screen tubes, the gas is well mixed and cooled before it encounters the superheater and hence the performance can be predicted more accurately. Due to the Lower log mean temperature difference and lower heat transfer coefficient, the surface area required will be greater and the device hence could therefore be more expensive than the radiant design.

16.25.4 Pendent Superheater

Pendent superheaters receive heat by both convection and radiation, they are normally hung from the top as shown in the Fig. 16.40, usually located in the crossover duct between the furnace and the back pass.

The outside tube diameters of the pendent superheaters normally falls in the range 32–51 mm and the tube thickness is usually in the range 3–7 mm.

16.25.5 Platen Superheater

These devices are made from flat panels of tubes located in the upper part of the furnace, where the gas temperature is high. The tubes of the platen superheater receive very high radiation as well as a heavy dust burden, so ultimate care should be taken while designing platen superheaters. The arrangement of platen superheater is shown in Fig. 16.41.

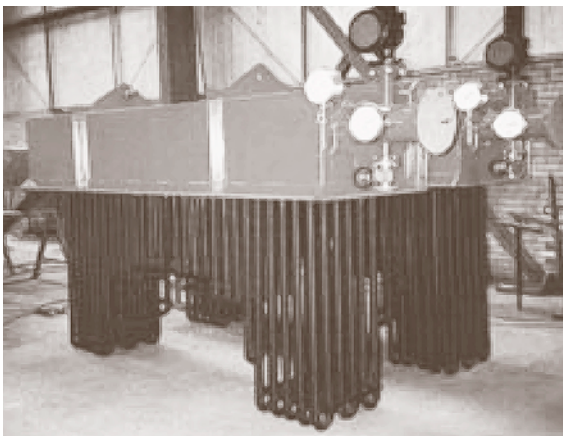


Fig. 16.40 Arrangement of a pendent superheater

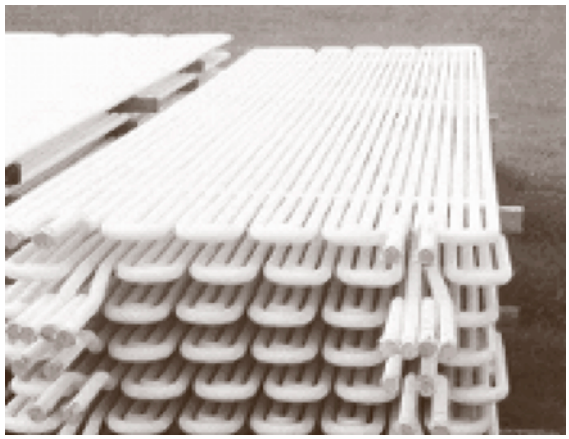


Fig. 16.41 Arrangement of a platen superheater

The mass flow velocity of steam in the platen superheater is normally in the range 800–1000 kg/(m^2 s). The outer diameter of the platen superheater tubes is in the range 32–42 mm. The number of parallel tubes in a platen is generally 15–35, depending on the design steam velocity.

16.25.6 Reheaters

The design considerations for reheaters are similar to those for superheaters. The reheater is usually located above the primary or convective superheater in the convective zone of utility boilers. A schematic view of a convective reheater is shown in Fig. 16.43. The pressure drop inside the reheater tubes has an important adverse effect on the efficiency of the turbine. The pressure drop through the reheater should be kept as low as possible. The tube diameter of the reheater is normally 42–60 mm and the overall heat transfer coefficient is 90–110 W/(m^2 K).

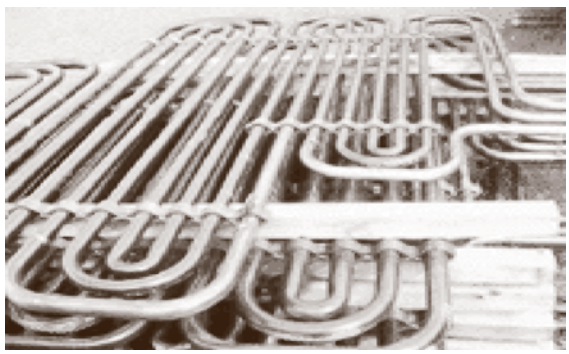


Fig. 16.42 Schematic view of convective reheaters

16.25.7 Economizers

The feedwater from the first high-pressure heater passing through a heat exchanger and heats up to the saturation temperature corresponding to the boiler pressure. This heat exchanger is normally called the economizer; it extracts waste heat from hot stack gases to heat the feedwater to the desired saturation temperature, hence the energy input to the boiler increases and the efficiency as well as economy of the power plant increase. The economizer is generally placed between the convective superheater and the air preheater.

Economizers are designed for downward flow of gas and upward flow of water, consisting of more than 250–300 coils in a staggered arrangement in a single bank. Figure 16.43 depicts the arrangement of the economizer.

Water enters from a lower header and flows through horizontal tubes that comprise the heating surface. Return bends at the ends of the tubing provide continuous tube elements whose upper ends connect to the outlet headers, which are in turn connected to the boiler drum by means of piping. Modern power plants use steel-tube-type economizers. The outside diameter of the economizer tubes is normally in the range 25–75 mm and the tube thickness is 3–5 mm; providing an extended surface over the economizer will enhance the heat transfer rate.

16.25.8 Feedwater Heaters

Feedwater heaters are used to raise the temperature of the water or to increase the mean temperature of heat addition in the cycle before it enters the boiler. The feedwater heater utilizes the steam which is extracted along the turbine expansion line for water heating. Feedwater heaters are used in a regenerative feedwater cycle to increase thermal efficiency and thus provide fuel savings.

Feedwater heaters are normally classified into two types

1. Open feedwater heaters
2. Closed feedwater heaters

Closed Feedwater Heaters

A closed feedwater heater is a shell-and-tube heat exchanger that warms feedwater by means of superheated steam or dry saturated or wet steam. Normally water flows inside the pipe and steam flows on the shell side. The arrangement of a closed feedwater heater in a typical steam turbine power plant is shown in Fig. 16.44.

According to the method of releasing the drain, closed feedwater arrangements are further classified into two types:

1. Drain cascaded forward
2. Drain cascaded backward

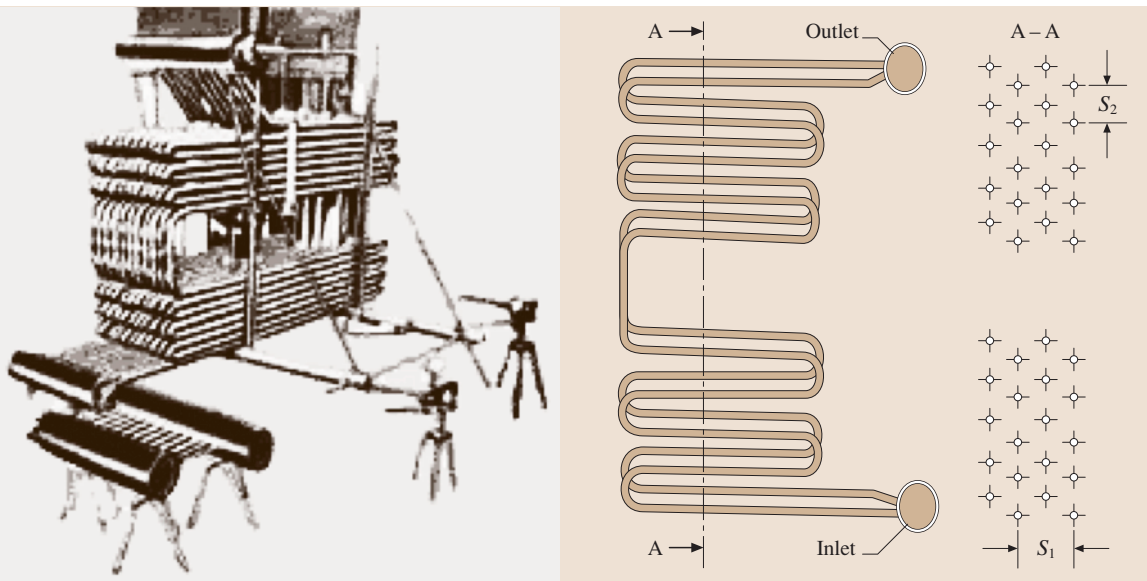


Fig. 16.43 Arrangement of an economizer

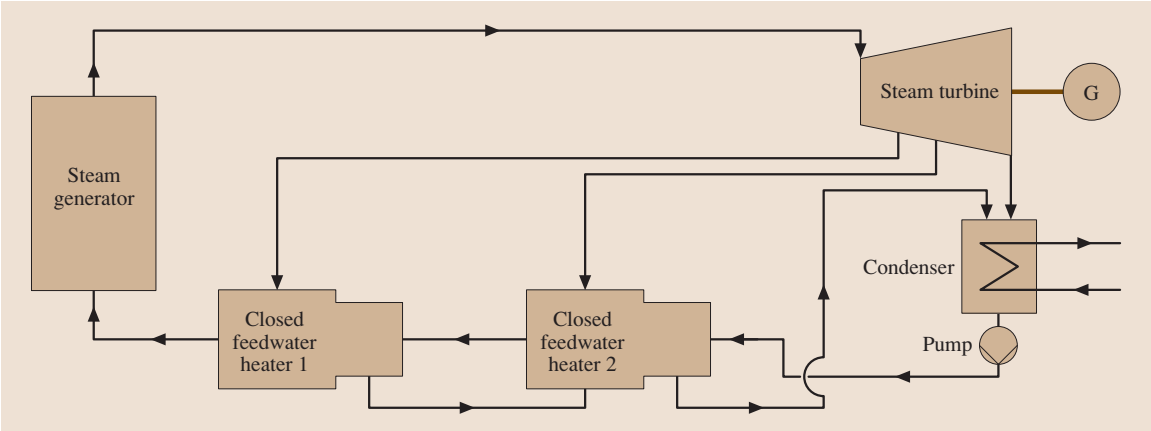


Fig. 16.44 Steam turbine regenerative cycle with closed feedwater heater (drains cascaded backward)

In the drain cascaded forward type the drain from the feedwater heater is mixed with the drain from the preceding feedwater heater. The reverse happens in the drain cascaded backward type, as shown in Fig. 16.44. Based on the state of the steam or level of turbine extraction feedwater heaters are further classified as high-, intermediate-, and low-pressure feedwater heaters. Modern power plants employing regeneration normally employ more than six feedwater heaters. There should always be one deaerator along with open feedwater heater. The purpose of the deaerator is to remove the dissolved oxygen presence in the system. High-pressure (HP) feedwater heaters utilize steam from the HP turbine and hence are called high-pressure feedwater heaters. A high-pressure closed feedwater heater has three zones, namely the desuperheating zone, the condensing zone, and the drain cooling zone. During the design of such a heat exchanger each zone is consider as a separate heat exchanger and the corresponding heat transfer coefficients and pressure drops are evaluated separately. The temperature duty diagram for a three-zone feedwater heater or high-pressure closed feedwater heater is shown in Fig. 16.45.

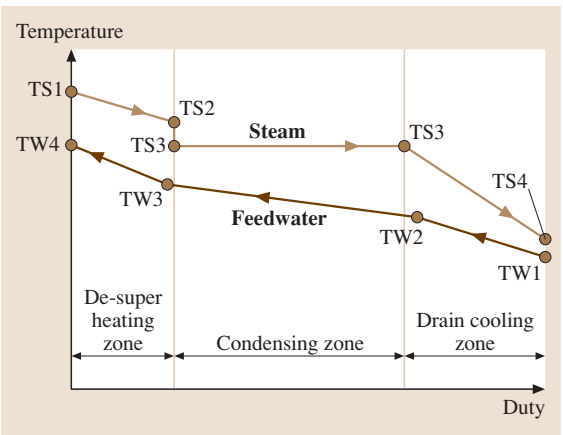


Fig. 16.45 Temperature–duty diagram for a three-zone feedwater heater

Open Feedwater Heaters

In open feedwater heaters, heat transfer takes place by direct mixing of steam and water. Normally open feedwater heaters are more efficient than closed feedwater heaters. Though the efficiency of open feedwater heater is higher, closed feedwater is normally used for modern power plants utilizing a large number of feedwater heaters to avoid a large number of pumps at each entrance and exit of the heater.

16.25.9 Air Preheaters

The flue gas leaving the chimney is normally in the range 280–480 °C. Leaving the flue to the atmosphere at this high temperature causes high energy losses. Air heaters are used to utilize this hot flue gas to heat the air required for combustion, and lead to an improvement in combustion efficiency. Since this is a gas-to-gas heat exchanger, its heat transfer surface area is extremely large. Recuperative and regenerative heaters are two different air preheaters normally employed in power plants.

16.25.10 Recuperative Air Preheater

A recuperative air preheater is nothing but a shell-and-tube heat exchanger in which hot flue gas flows inside

the tubes and air flows outside. Since this is a gas-to-gas heat exchanger it requires a huge heat transfer surface area and hence larger size.

16.25.11 Rotary or Regenerative Air Preheater

Rotary preheaters works on the counterflow principle, and consist of a rotor and housing. The rotor is normally divided into 12–24 radial divisions of heat transfer elements and is made up of steel sheets. The rotor is driven by an electric motor and is coupled with worm-gear drive that helps to reduce the speed of the rotor device to 2–6 rpm. During the rotation through the flue gas side the heat transfer element absorbs heat which is later given off during the rotation through the air section. Based on the number of sections rotary preheaters are further classified into three types: bisector, trisector, and quadsector types. Trisector-type air preheaters are divided into three sections: one for the flue gas, one for the primary, and one for the secondary section. In the quadsector type, the secondary air section is divided into two sections, taking up primary air. Control of the leakage of air into the flue gases is very important to avoid energy loss in the system. Various types of sealing systems are normally employed:

1. Radial sealing system
2. Axial sealing system
3. Circumferential sealing system
4. Shaft sealing system

16.26 Energy Balance Analysis of a Furnace/Combustion System

16.26.1 Performance Analysis of a Furnace

The following procedures should be adopted to carry out the performance analysis of the furnace:

1. Obtain the ultimate fuel analysis
2. Compute the equivalent chemical formula
3. Select the recommended exhaust gas composition
4. Write and balance the combustion equation

$$\begin{aligned}&\mathrm{C}_X\mathrm{H}_Y\mathrm{S}_Z\mathrm{O}_K + \varepsilon 4.76\left(X+\frac{Y}{4}+Z-\frac{K}{2}\right)\text{ air}\\&\quad+\text{moisture in air}+\text{ash moisture in fuel}\\&\rightarrow P\mathrm{CO}_2+Q\mathrm{H}_2\mathrm{O}+R\mathrm{SO}_2+T\mathrm{N}_2+U\mathrm{O}_2\\&\quad+V\mathrm{CO}+W\mathrm{C}+\text{ash}\,.\end{aligned}$$

Dry Exhaust Gases.

$$P\mathrm{CO}_2+R\mathrm{SO}_2+T\mathrm{N}_2+U\mathrm{O}_2+V\mathrm{CO}\quad\text{(kmol)}\,.$$

1. Carry out first law analysis to calculate the theoretical combustion temperature.
2. Calculate the total number of moles of wet exhaust gas for 100 kg of fuel

$$n_{\text{ex.gas}} = P+Q+R+T+U+V\,.$$

3. $100\,\text{CV of fuel} = n_{\text{ex.gas}}c_{\text{p}_{\text{exhaust gas}}}(T_{\text{th}}-T_{\text{atm}})\,.$
4. Calculate the total heat transfer area of the furnace A_{furnace}.

16.26.2 Analysis

$$\begin{aligned}&\mathrm{C}_X\mathrm{H}_Y\mathrm{S}_Z\mathrm{O}_K + \varepsilon 4.76\left(X+\frac{Y}{4}+Z-\frac{K}{2}\right)\text{ air}\\&\quad+\text{moisture in air}+\text{ash moisture in fuel}\\&\rightarrow P\mathrm{CO}_2+Q\mathrm{H}_2\mathrm{O}+R\mathrm{SO}_2+T\mathrm{N}_2+U\mathrm{O}_2\\&\quad+V\mathrm{CO}+W\mathrm{C}+\text{ash}\,.\end{aligned}$$

Dry exhaust gases

$$P\mathrm{CO}_2+R\mathrm{SO}_2+T\mathrm{N}_2+U\mathrm{O}_2+V\mathrm{CO}\quad\text{(kmol)}\,.$$

The volume of gases is directly proportional to the number of moles

Volume fraction = mole fraction

Volume fraction of CO_2

$$x_1 = P\frac{100}{(P+R+T+U+V)}\,.$$

Volume fraction of CO

$$x_2 = V\mathrm{CO}\frac{100}{(P+R+T+U+V)}\,.$$

Volume fraction of SO_2

$$x_3 = R\frac{100}{(P+R+T+U+V)}\,.$$

Volume fraction of O_2

$$x_4 = U\frac{100}{(P+R+T+U+V)} .$$

Volume fraction of N_2

$$x_5 = T\frac{100}{(P+R+T+U+V)} .$$

These are dry gas volume fractions. Emission measurement devices only indicate dry gas volume fractions.

Measurements

Volume flow rate of air, volume flow rate of exhaust are obtained from exhaust gas analysis

$$x_1 + x_2 + x_3 + x_4 + x_5 = 100 \text{ or } 1 .$$

Ultimate Analysis of Coal

$$n\mathrm{C}_X\mathrm{H}_Y\mathrm{S}_Z\mathrm{O}_K + \varepsilon n 4.76\left(X + \frac{Y}{4} + Z - \frac{K}{2}\right) \text{ air}$$
$$+ \text{moisture in air} + \text{ash and moisture in fuel}$$
$$\rightarrow x_1\mathrm{CO}_2 + x_6\mathrm{H}_2\mathrm{O} + x_3\mathrm{SO}_2 + x_5\mathrm{N}_2 + x_4\mathrm{O}_2$$
$$+ x_2\mathrm{CO} + x_7\mathrm{C} + \text{ash} ,$$

where x_1, x_2, x_3, x_4, and x_5 are dry volume fractions or percentages.

Conservation Species. Conservation of carbon

$$nX = x_1 + x_2 + x_7 .$$

Conservation of hydrogen

$$nY = 2x_6 .$$

Conservation of oxygen

$$nK + 2n\varepsilon\left(X + \frac{Y}{4} + Z - \frac{K}{2}\right)$$
$$= 2x_1 + x_2 + 2x_3 + 2x_4 + x_6 .$$

Conservation of nitrogen

$$\varepsilon n 3.76\left(X + \frac{Y}{4} + Z - \frac{K}{2}\right) = x_5 .$$

Conservation of sulfur

$$nZ = x_3 .$$

By rearranging the terms, we obtain

$$\mathrm{C}_X\mathrm{H}_Y\mathrm{S}_Z\mathrm{O}_K + \varepsilon 4.76\left(X + \frac{Y}{4} + Z - \frac{K}{2}\right) \text{ air}$$
$$+ \text{moisture in air} + \text{ash moisture in fuel}$$
$$\rightarrow P\mathrm{CO}_2 + Q\mathrm{H}_2\mathrm{O} + R\mathrm{SO}_2 + T\mathrm{N}_2 + U\mathrm{O}_2$$
$$+ V\mathrm{CO} + W\mathrm{C} + \text{ash} .$$

16.26.3 First Law Analysis of Combustion

$$\mathrm{C}_X\mathrm{H}_Y\mathrm{S}_Z\mathrm{O}_K + \varepsilon 4.76\left(X + \frac{Y}{4} + Z - \frac{K}{2}\right) \text{ air}$$
$$+ \text{moisture in air} + \text{ash} + \text{moisture in fuel}$$
$$\rightarrow P\mathrm{CO}_2 + Q\mathrm{H}_2\mathrm{O} + R\mathrm{SO}_2 + T\mathrm{N}_2 + U\mathrm{O}_2$$
$$+ V\mathrm{CO} + W\mathrm{C} + \text{ash} ,$$

$$\sum Q_{\mathrm{CV}} + n_{\mathrm{air}}h_{\mathrm{f,air}} + n_{\mathrm{fuel}}h_{\mathrm{f,fuel}}$$
$$= \sum_{i=1}^{n} n_i h_{\mathrm{f,fluegas},i} + \sum W_{\mathrm{CV}} . \qquad (16.63)$$

Furnace characterization criteria

$$A_{\mathrm{furnace}} = \frac{Gm_{\mathrm{f}}c_{\mathrm{P}}}{T_{\mathrm{th}}^3}\left[\frac{1}{m}\left(\frac{T_{\mathrm{th}}}{T_{\mathrm{out}}} - 1\right)\right]^{1/0.6} , \qquad (16.64)$$

where G is the furnace quality factor, M is the temperature field coefficient, T_{th} is the theoretical combustion temperature, A_{furnace} is the total surface area of the furnace, and m_{f} is the mass flow rate of fuel.

16.26.4 Boiler Fuel Consumption and Efficiency Calculation

For any fuel there is a minimum quantity of oxygen required for complete combustion. The amount of air that contains this minimum quantity of oxygen is called the *theoretical air*; it is the measure of capability of the boiler to transfer heat liberated in the furnace to water and steam. The boiler efficiency may be expressed in any one of the following methods.

16.26.5 Various Energy Losses in a Steam Generator

1. Heat loss from the furnace surface area
2. Unburned carbon loss
3. Incomplete combustion loss
4. Loss due to hot ash
5. Loss due to moisture in the air
6. Loss due to moisture in the fuel
7. Loss due to combustion-generated moisture
8. Dry exhaust gas losses

The pictorial representation of the Shanky diagram, as shown in Fig. 16.47, represents the various energy losses that take place in a steam generator.

The first law analysis steam generator in steady state steady flow (SSSF) mode in molar form (see Fig. 16.48)

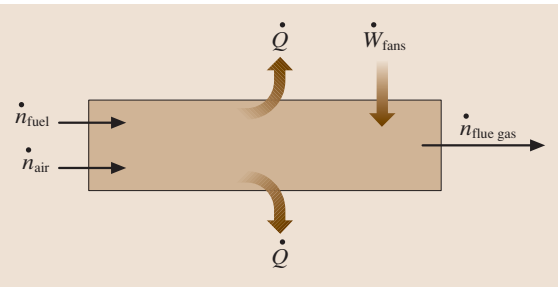


Fig. 16.46 Energy balance diagram

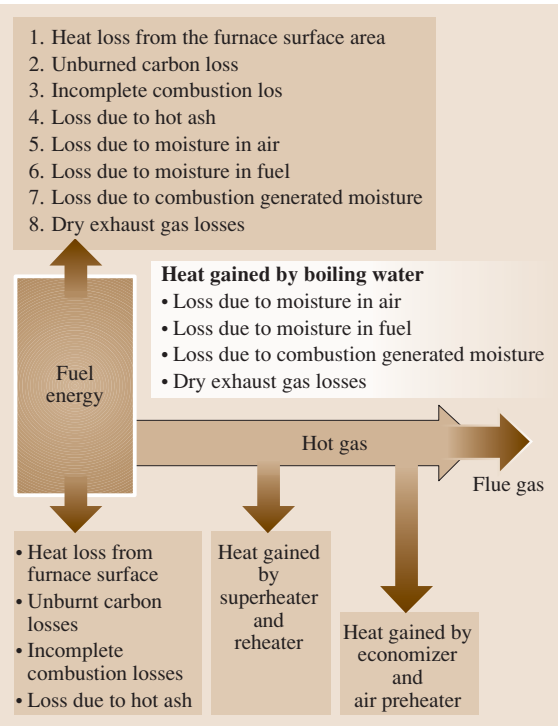


Fig. 16.47 Shanky diagram showing various losses in the steam generator

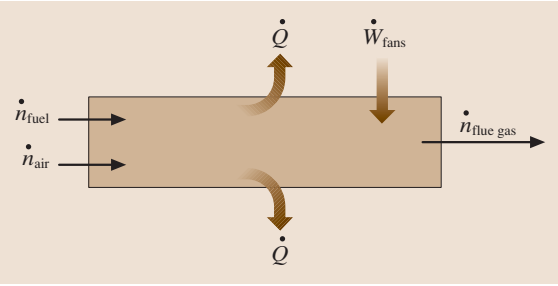


Fig. 16.48 Furnace energy balance

yields

$$\sum \dot{Q} + \dot{n}_{\text{air}} h_{\text{air}} + \dot{n}_{\text{fuel}} h_{\text{fuel}} = \sum \dot{n}_{\text{fluegas}} h_{\text{fluegas}} + \sum \dot{w} \, .$$

Dry Exhaust Gas Losses (DEGL)

For 100 kg of fuel

$$Q_{\text{DEGL}} = \Delta n_{\text{fluegas}} \Delta h_{\text{fluegas}} \, ,$$

$$Q_{\text{DEGL}} = n_{\text{CO}_2} \Delta h_{\text{CO}_2} + n_{\text{CO}} \Delta h_{\text{CO}} + n_{\text{O}_2} \Delta h_{\text{O}_2} + n_{\text{N}_2} \Delta h_{\text{N}_2} + n_{\text{SO}_2} \Delta h_{\text{SO}_2} \text{ kJ} \, ,$$

$$Q_{\text{DEGL}} = P \Delta h_{\text{CO}_2} + R \Delta h_{\text{SO}_2} + T \Delta h_{\text{N}_2} + U \Delta h_{\text{O}_2} + v \Delta h_{\text{CO}} \text{ kJ} \, .$$

Alternate method. The total number of moles of dry exhaust gas $n_{\text{ex.gas}} = P + R + T + U + V$ is

$$Q_{\text{DEGL}} = n_{\text{ex.gas}} C_{\text{p,ex.gas}} (T_{\text{ex.gas}} - T_{\text{atm}}) \, .$$

Accurate calculation of the gas enthalpy gives, for any gas

$$\mathrm{d}h = c_{\text{p}} \, \mathrm{d}T \, ,$$

$$\Delta h = \int_{\text{ambient}}^{\text{SG}_{\text{exit}}} \mathrm{d}h = \int_{T_{\text{amb}}}^{T_{\text{exit}}} c_{\text{p}}(T) \, \mathrm{d}T \, . \qquad (16.65)$$

Unburnt Carbon Loss (UCL)

For 100 kg of fuel

$$Q_{\text{UCL}} = \text{MC } W \text{ calorific value of carbon} \quad \text{(kJ)} \, ,$$

$$Q_{\text{UCL}} = 12 W \, 33\,820 \text{ kJ} \, , \qquad (16.66)$$

where MC is the molecular weight of carbon.

Incomplete Combustion Loss (ICL)

For 100 kg of fuel

$$Q_{\text{ICL}} = \text{MCO} VCV \text{ of CO} \quad \text{(kJ)} \, ,$$

$$Q_{\text{ICL}} = 28 V \, 23\,717 \text{ kJ} \, . \qquad (16.67)$$

Losses due to Moisture in the Combustion Air (MCAL)

$$Q_{\text{MCAL}} = e 4.76 \left(X + \frac{Y}{2} + Z - \frac{K}{2} \right) \times 29.9 w c_{\text{steam}} (T_{\text{g}} - 25) \quad \text{(kJ)} \, , \qquad (16.68)$$

where w is the absolute or specific humidity (kg of moisture per kg of dry air), c_{steam} is the specific heat

of steam at constant pressure (1.88 kJ/(kg K)), and T_g is the temperature of exhaust gas.

Losses due to Moisture in Fuel and Combustion Generated Moisture

For 100 kg of fuel

$$Q_{ML} = (M + 9Y)[2442 + c_{steam}(T_g - 25)] \quad \text{(kJ)}\,, \tag{16.69}$$

where M is the percentage moisture content in the fuel and Y is the combustible hydrogen atoms in the fuel.

Losses due to Hot Ash or Slag (ASL)

For 100 kg of fuel

$$Q_{ASL} = A c_{p,as} T_{ash}\,, \tag{16.70}$$

where $c_{p,ash}$ is the specific heat of ash (0.55–0.6 kJ/(kg, K)), T_{ash} is the temperature of the ash or slag, and T_{ash} varies from 300 to 800 °C.

Radiation and Unaccounted-for Losses (RUL)

This calculation captures losses due to radiation and incomplete combustion resulting in hydrogen and hydrocarbons in the flue gases. While the radiation and unaccounted loss is relatively small, it is difficult to determine accurately. In practice, this loss is 3–5%.

$$Q_{RCL} = A_s(h_s)(T_{surface} - T_{amb}) \quad \text{(kW)}\,, \tag{16.71}$$

where A_s is the total surface area (m^2), and h_s is the surface heat transfer coefficient.

16.27 Performance of Steam Generator

16.27.1 Boiler Efficiency

This is the measure of the capability of the boiler to transfer heat liberated in the furnace into water and steam. The boiler efficiency may be expressed in any of the following methods

$$\eta_{boiler} = \frac{\text{mass flow rate of steam} \times (\text{steam heat} - \text{feedwater heat})}{\text{fuel mass} \times \text{heating value of fuel}}$$

or

$$\eta_{boiler} = \frac{(\text{HHV} - \text{total loss})}{\text{HHV}}\,, \tag{16.72}$$

where HHV is the higher heating value of the fuel.

16.28 Furnace Design

There are two aspects of furnace design. The first is concerned with the generation of the heat; the second part involves the absorption of the heat in the furnace. The amount of fuel can be burned in the given furnace volume, liberating the required amount of heat.

The heat release rate and furnace gas temperature are two of the important parameters used for the design of the size of the furnace. The heat release rate is expressed on three different bases: furnace volume (q_v), furnace cross-sectional area, and water wall area in the burner region (q_b).

The important thermal characteristic of the furnace for design analysis are:

- Heat release rate per unit cross-sectional area
- Heat release rate per unit volume
- Heat release rate per unit wall area of the burner region

16.28.1 Heat Release Rate per Unit Volume q_v

The amount of heat generated by the combustion of fuel in a unit effective volume of the furnace is given by

$$q_v = \frac{m_f \times \text{LHV}}{V_{furnace}} \quad \text{(kW/m}^3\text{)}\,, \tag{16.73}$$

where m_f is the designed fuel consumption rate (kg/s), LHV is the lower heating value of the fuel (kJ/kg), V_f is the volume of the furnace ($a \times b \times h_f$), and h_f is the height of the furnace.

The value of q_v depends on the coal type and type of furnace.

The volumetric heat release rate also depends on the ash characteristic, firing method, and the arrangement of the burners. The proper selection of the volumetric flow

rate in accord with the heat release rate will ensure that fuel particles are substantially burned.

16.28.2 Heat Release Rate per Unit Wall Area of the Burner Region

One of the most important regions in the furnace is the burner region. The heat release rate in the burner region is calculated on the basis of the water wall in this region. The heat release rate per unit wall area of the furnace depends on the following parameters [16.16]:

1. Ash characteristic
2. Fuel ignition characteristics
3. Firing method
4. Arrangement of burners

The heat release rate per unit wall area of the burner region my be written

$$q_b = \frac{m_f \mathrm{LHV}}{2(a+b)h_b} \quad (\mathrm{kW/m^2}) , \tag{16.74}$$

where a and b are the width and depth of the furnace, respectively, and h_b is the distance between the top edge of the uppermost burner and the lower edge of the lowest burner.

16.28.3 Heat Release Rate per Unit Cross-Sectional Area

This is the amount of heat released per unit cross section of the furnace. It is given by

$$q_f = \frac{m_f \mathrm{LHV}}{A_{\mathrm{furnace}}} \quad (\mathrm{kW/m^2}) , \tag{16.75}$$

where A_{furnace} is the cross-sectional area of the furnace in $\mathrm{m^2}$.

16.28.4 Furnace Exit Gas Temperature

The furnace exit gas temperature is an important design parameter. It determines the rate of heat absorption by the radiant heating surface in the furnace and that by the convective heating surface of the furnace. The optimum value of the furnace exit gas temperature is 1200–1400 °C.

16.28.5 Example Problem

An optimal operation test on a model steam generator gives the following information:

- Ultimate analysis: C: 63.4%, H: 5.7%, O: 16.8%, N: 10%, ash: 8.9%, moisture: 13%
- HHV of coal: 33 318 kJ/kg
- Combustible solid refuse: 7.5%
- Dry exhaust gas analysis: CO_2: 15.4%, CO: 0.5%, O_2: 2.8%, N_2: 81.3%
- Ambient conditions: 50 °C and 100 kPa
- Temperature of air entering the furnace: 235 °C

Design a PC (pulverized coal) furnace for a steam generator with a thermal capacity of 1000 MW with the following characteristics:

- Steam generator efficiency of 0.86
- Furnace quality factor of 0.406×10^8
- Temperature field coefficient of $M = 0.405$
- Thermal capacity of the gas of $c_p = 1.17\,\mathrm{kJ/(kg\,K)}$

A furnace can be characterized geometrically by its linear dimensions: front width a, the depth b, and the height h_f (Fig. 16.29), which are estimated according to the rated fuel consumption and the thermal, physical, and chemical properties of the fuel to be used.

Flow Rate of Fuel

$$\begin{aligned} Q_{\mathrm{boiler}} &= m_{\mathrm{coal}} \mathrm{HHV} \eta_{\mathrm{SG}} , \\ 1\,000\,000 &= m_{\mathrm{coal}} 33\,318 \eta_{\mathrm{SG}} , \\ m_{\mathrm{coal}} &= 34.899\,\mathrm{kg/s} ; \\ \mathrm{LHV} &= \mathrm{HHV} - \frac{m_{\mathrm{H_2O}}}{\mathrm{kg\ of\ fuel}} 2442 , \\ \mathrm{LHV} &= 323\,827\,\mathrm{kJ/kg} . \end{aligned}$$

Here, η_{SG} is the efficiency of the steam generator.

Heat Release Rate per Unit Volume

$$q_v = \frac{m_f \mathrm{LHV}}{V_f} \quad (\mathrm{kW/m^3}) . \tag{16.76}$$

The large content of H and O is largely volatile matter and hence the given composition is bituminous. For bituminous coal the range of value of q_v is $0.14–0.20\,\mathrm{MW/m^3}$ [16.16].

Substituting the values of q_v, m_f, and LHV (Table 16.1) we find the volume of the furnace to be

$$V = 7532.05\,\mathrm{m^3} .$$

Heat Release Rate per Unit Cross-Sectional Area

This is the amount of heat released per unit cross section of the furnace. It is given by

$$q_a = \frac{m_f \mathrm{LHV}}{A_{\mathrm{grade}}} \quad (\mathrm{kW/m^2}) , \tag{16.77}$$

Table 16.1 Typical values of the volumetric heat release (q_v) in MW/m^3

Coal type	Dry-bottom furnace q_v (MW/m^3)	Wet (slagging) bottom furnace q_v (MW/m^3)		
		Open furnace	Half-open furnace	Slagging pool
Anthracite	0.110–0.140	≤ 0.145	≤ 0.169	0.523–0.598
Semi-anthracite	0.116–0.163	0.151–0.186	0.163–0.198	0.523–0.698
Bituminous	0.14–0.20	–	–	–
Oil	0.23–0.35	–	–	–
Lignite	0.09–0.15	≤ 0.186	≤ 0.198	0.523–0.640
Gas	0.35	–	–	–

Table 16.2 Upper limits of q_a for tangentially fired furnaces

Boiler capacity (t/h)	Upper limit of q_a (MW/m^2)		
	ST[a] ≤ 1300 °C	ST = 1300 °C	ST ≥ 1300 °C
130	2.13	2.56	2.59
220	2.79	3.37	3.91
420	3.65	4.49	5.12
500	3.91	4.65	5.44
1000	4.42	5.12	6.16
1500	4.77	5.45	6.63

[a] ST = softening temperature of ash (°C)

where A_{grade} is the cross-sectional area of the grade in m^2

$$A_{\text{grade}} = ab = \frac{m_f \text{LHV}}{q_a} \quad (\text{kW/m}^2)\,. \tag{16.78}$$

Substitute the value of q_a, m_f, and LHV (Table 16.2) in (16.78) we find the grade area $ab = 441.46\,\text{m}^2$.

Heat Release Rate per Unit Wall Area of the Burner Region

The heat release rate per unit wall area of the burner region may be written

$$q_b = \frac{m_f \text{LHV}}{2(a+b)h_b} \quad (\text{kW/m}^2)\,. \tag{16.79}$$

The recommended value of the burner region heat release rate was taken as

$$h_b = 1\,\text{MW/m}^3\,.$$

$$2(a+b)h_b = 1129.80\,\text{m}^2\,.$$

Based on Tables 16.1 and 16.3, corresponding to the boiler capacity, the minimum width was chosen as $b_{\min} = 6\,\text{m}$ and $h_{\text{furnace,min}} = 11\,\text{m}$.

Based on the above constraints suitable values for a and b are

$$a = 20.62\,\text{m}\,,$$

$$b = 21.4\,\text{m}\,,$$

$$h_b = \frac{1129.8}{2(21.4+20.62)} = 13.44\,\text{m}\,.$$

The volume of the furnace region (Fig. 16.29) is therefore

$$V_{\text{furnace}} = h_{\text{furnace}}ab - \frac{a}{2}(d+d+d\tan\beta+d\tan\alpha)d\,, \tag{16.80}$$

$$V_{\text{furnace}} = h_{\text{furnace}}ab - \frac{ad}{2}(2d+d\tan\beta+d\tan\alpha)\,, \tag{16.81}$$

$$V_{\text{furnace}} = h_{\text{furnace}}ab - \frac{ad^2}{2}(2+\tan\beta+\tan\alpha)\,. \tag{16.82}$$

Substituting the values of a, b, α, β, and V_{furnace} into (16.82) yields

$$h_{\text{furnace}} = 19.33\,\text{m}\,.$$

To find the height of the hopper we insert the data into

$$h_h = \left(\frac{b-e}{2}\right)\tan\gamma = 14.56\,\text{m}\,. \tag{16.83}$$

From the geometry we calculate the total surface area as

$$\begin{aligned} &a(h_f+h_b)+a(h_f+h_b-d-d\tan\alpha-d\tan\beta \\ &\quad +d\sec\alpha+d\sec\beta)+2b(h_f+h_b) \\ &\quad -d(d+d+d\tan\alpha+d\tan\beta)+2\left[\frac{1}{2}h_h(b+e)\right] \\ &\quad +2ah_h\,\text{cos ec}\,\gamma\,. \end{aligned} \tag{16.84}$$

Table 16.3 Lower limit of h_{furnace} (m)

Boiler capacity (t/h)	65–75	130	220	420	670
Anthracite	8	11	13	17	18
Bituminous	7	9	12	14	17

Substituting all the values into (16.84), the surface area of the furnace is calculated to be 3793.98 m^2.

Adiabatic Flame Temperature

Ultimate Analysis.

$$\text{C}: \quad \frac{64.4}{12} = 5.283$$

$$\text{H}: \quad 5.7 - \frac{13}{9} = 4.255$$

$$\text{O}: \quad \frac{16.8}{16} - \left[\left(\frac{8}{9}\right)\left(\frac{13}{16}\right)\right] = 0.327$$

$$\text{N}: \quad \frac{1}{14} = 0.071$$

$$\text{Ash} \quad = 8.9$$

$$\text{Moisture} \quad = \frac{13}{18}$$

$$\text{C}_X\text{H}_Y\text{S}_Z\text{O}_K + \left(X + \frac{Y}{4} + Z - \frac{K}{2}\right)\text{O}_2$$
$$\rightarrow 15.4\text{CO}_2 + 0.5\text{H}_2\text{O} + 2.8\text{O}_2 + 81.3\text{N}_2 + Z\text{H}_2\text{O}$$

For 100 kg of fuel

$$\text{C}_{5.28}\text{H}_{4.25}\text{O}_{0.327} + 4.76 \cdot 1.162 \cdot 6.152\text{air}$$
$$\rightarrow 5.11\text{CO}_2 + 0.166\text{H}_2\text{O} + 0.933\text{O}_2$$
$$+ 27.1\text{N}_2 + 2.13\text{H}_2\text{O}\,,$$

n_{exit} (mole of exit gases)

$$= 5.11 + 0.166 + 0.93 + 27.1 + 2.13 = 35.43\,,$$
$$100 \cdot 33\,318 = 35.43 \cdot 40 \cdot (T_{\text{th}} - T_{\text{atm}})\,,$$

$$T_{\text{th}} = 2624\,\text{K}\,.$$

The ash softening temperature is $\geq 1250\,^{\circ}\text{C}$ and $T_{\text{out}} \leq 1250\,^{\circ}\text{C} = 1523$ K.

100 kg of fuel generates 1056.71 of exhaust flue gas with m_{gas} for a 34.89 kg/s of flow rate of fuel, $m_{\text{gas}} = 368.68$ kg/s.

To find T_{out} we use

$$A_{\text{furnace}} = \frac{G m_{\text{f}} c_{\text{p}}}{T_{\text{th}}^3}\left[\frac{1}{m}\left(\frac{T_{\text{th}}}{T_{\text{out}}} - 1\right)\right]^{1/0.6}\,, \quad (16.85)$$

where G is the furnace quality factor, M is the temperature field coefficient, T_{th} is the theoretical combustion temperature, A_{furnace} is the total surface area of furnace, and m_{f} is the mass flow rate of fuel.

Substituting all these values in (16.85) we find $T_{\text{out}} = 1365.95$ K.

T_{out} is < 1523 K (ash softening temperature), so the design is safe.

16.29 Strength Calculations

Special care must be taken in the design and stress analysis of steam generators because of the application of high pressure and temperature involved in the system. Allowable stresses in the pressure vessel depend on the nature of the loading in the pressure vessel and the response to this loading.

Stress can be classified into:

1. Primary stress
2. Secondary stress
3. Peak stress

The *primary stress* is developed by the mechanical load; it can cause mechanical failure of the vessel. An example of this kind of stress is that produced by internal pressure such as in a steam drum. *Secondary stress* is due to mechanical load or thermal expansion. *Peak stress* is concentrated in highly localized area at abrupt geometry changes.

16.29.1 Mathematical Formulae for Stress

The basic equation for the longitudinal stress σ_1 and hoop stress σ_2 in a vessel of thickness of h, longitudinal radius r_1, and circumferential stress r_2, which is subjected to a pressure p is given by

$$\frac{\sigma_1}{r_1} + \frac{\sigma_2}{r_2} = \frac{p}{h}\,. \quad (16.86)$$

From this equation, and by equating the total pressure load with the longitudinal forces acting on a transverse section of this vessel, the stresses in the commonly used shells of revolution can be found.

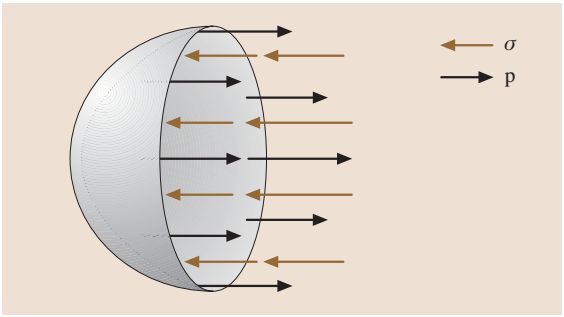


Fig. 16.49 Cross-sectional view of spherical vessel

16.29.2 Stress Analysis Methods

Stress analysis of pressure vessel can be performed by analytical or experimental methods. The general shapes of a pressure vessel are spheres, cylinders, and ellipses.

Spherical Vessel ($r_1 = r_2 = r$). Consider a spherical pressure vessel with radius r and wall thickness h subjected to an internal pressure. All four normal stresses on a small stress in the wall must be identical, due to symmetry. Furthermore there can be no shear stress. The normal stresses σ can be related to the pressure p by inspecting the free-body diagram of the pressure vessel. To simplify the analysis, cut the vessel in half as illustrated in Fig. 16.49.

The stress around the wall must have a net resultant to balance the internal pressure across the cross section

$$\sigma h 2\pi r = p\pi r^2 ,$$

$$\sigma_1 = \frac{pr}{2h} , \tag{16.87}$$

$$\sigma_2 = \frac{pr}{2h} . \tag{16.88}$$

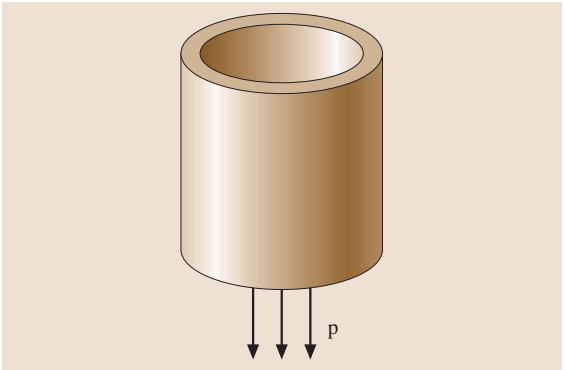


Fig. 16.50 Stress analysis of cylindrical vessel

Cylindrical Vessel ($r_1 = \infty$, $r_2 = r$).

$$\sigma_1 = \frac{pr}{2h} , \tag{16.89}$$

$$\sigma_2 = \frac{pr}{h} . \tag{16.90}$$

Conical Vessel $r_1 = \infty$, $r_2 = r/\cos\alpha$, Where α is the Half-Cone Angle.

$$\sigma_1 = \frac{pr}{2h\cos\alpha} , \tag{16.91}$$

$$\sigma_2 = \frac{pr}{h\cos\alpha} . \tag{16.92}$$

Elliptical Vessel.

$$\sigma_1 = \frac{pr_2}{2h} , \tag{16.93}$$

$$\sigma_2 = \frac{p}{h}\left(r_2 - \frac{r_2^2}{2r_1}\right) . \tag{16.94}$$

16.29.3 Design Pressure and Temperature

Generally, the design pressure is the maximum allowable working pressure. It should not be less than the highest set pressure of any safety valve. The determination of the allowable stresses is based upon the design temperature T, which should not be taken to be less than the mean wall metal temperature (through thickness) expected under operating conditions for the part considered. The design temperature is to be stated by the manufacturer on the drawings of the pressure parts submitted for consideration.

The design of pressure parts is based on the allowable stress S in $\mathrm{N/mm^2}$.

The minimum thickness of straight tubes is to be determined as

$$t = \frac{pD_o}{20SE + p} + C , \tag{16.95}$$

where p is the design pressure (bar), t is the minimum thickness (mm), D_o is the outside diameter (mm), S is the allowable stress ($\mathrm{N/mm^2}$), E is the weld efficiency of longitudinally welded tubes, and C is the corrosion allowance (mm).

The efficiency factor E is the welding efficiency of the longitudinal joint or of ligaments between tube holes or other openings.

16.30 Heat Transfer Calculation

16.30.1 Heat Exchangers

Heat exchangers play a vital role in power plant to transfer heat from hot to cold fluids. Heat exchangers with different flow configurations such as parallel flow, counterflow, and crossflow types are generally used.

Heat Transfer Analysis in a Counterflow Heat Exchanger

A counterflow heat exchanger, where the fluid moves in parallel but opposite direction, is shown in Fig. 16.51.

The thermal design of a heat changer involves the calculation of the surface area required to transfer heat at a given rate for given flow rates and fluid temperatures. The size of the heat exchanger can be obtained from the general heat transfer equation,

$$Q = U_o A_o \Delta T_{lm} \,, \tag{16.96}$$

where A_o is the outside heat transfer surface area based on the outside diameter of the tube, U_o is the overall heat transfer coefficient based on the outside diameter of the tube, and ΔT_{lm} is the log mean temperature difference

$$\Delta T_{lm} = \frac{\Delta T_1 - \Delta T_2}{\ln(\Delta T_1 / \Delta T_2)} \,, \tag{16.97}$$

where $\Delta T_1 = T_{h1} - T_{c2}$ and $\Delta T_2 = T_{h2} - T_{c1}$ for a counterflow heat exchanger.

16.30.2 Flow Resistance

Friction resistance values for the actual pipes and volume flows may be obtained from special charts made for the pipes or tubes considered.

Minor pressure losses due to fittings as bends, elbows, valves, and similar may be calculated as

$$p_2 = \xi \frac{\rho v^2}{2} \text{ or expressed as } \textit{head} \tag{16.98}$$

$$h_{loss} = \xi \frac{v^2}{2} g \,, \tag{16.99}$$

where ξ is the minor loss coefficient, p_{loss} is the pressure loss (Pa = N/m^2), ρ is the density (kg/m^3), v is the flow velocity (m/s), h_{loss} is the head loss (m), and g is the acceleration of gravity (m/s^2).

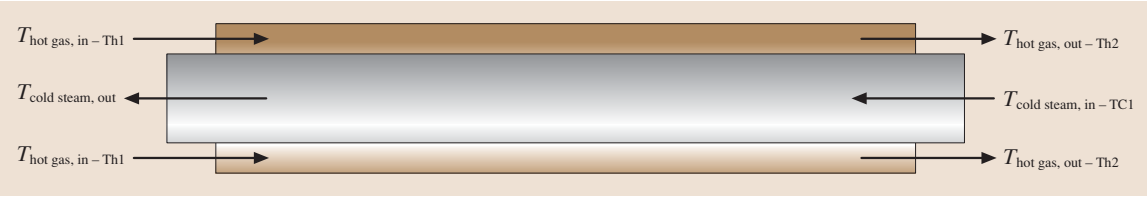


Fig. 16.51 Heat transfer along a heat exchanger

16.31 Nuclear Reactors

Nuclear reactors are devices designed to maintain a chain reaction producing neutrons generated by the fission of heavy nuclei. Nuclear power plants utilizing power reactors are dedicated to generate heat, mainly for electricity production. They are operated in more than 50 countries [16.17].

16.31.1 Components of a Nuclear Reactor

The major components of a nuclear reactor are:

1. The fuel core
2. The moderator and coolant
3. The control rods
4. The reactor vessel

Fuel Core

The fuel core contains the nuclear fuel and is the part of the reactor where the fission reaction takes place. The nuclear fuel may be either natural or enriched uranium. Natural uranium contains 0.71% fissile ^{235}U and 99.28% fertile ^{238}U and fertile thorium Th-232. The enriched uranium is produced in a gaseous diffusion process and is expected to have a ^{235}U content of 2–33%.

The nuclear fuel is generally contained in cylindrical rods surrounded by cladding materials. The fuel-rod

cladding materials must not only be able to maintain the shape of the fuel rod, but also to sustain the reactor conditions. Materials used for these components include aluminum, magnesium, stainless steel, and graphite.

Moderator and Coolant

The moderator is the substance used in a nuclear reactor to reduce the energy of fast neutrons to thermal neutrons. Liquid and solid materials with low atomic mass number and low neutron capture cross section should be suitable. These include light water, heavy water, carbon, and beryllium.

The reactor coolant is used to remove heat from the reactor fuel core. The conditions for a better coolant include high specific heat, high thermal conductivity, and high boiling point at low pressure. The coolant should also have low power demand for pumping, low cost, and a high degree of stability in the reactor environment.

Control Rods

Control rods are used to slow down or speed up a chain reaction. Elements like boron and cadmium are used in a control rod to absorb fast neutrons and thereby control the chain reaction. An automatic retractable mechanism helps to insert the control rods into the fuel core or withdraw them to slow down or speed up the chain reaction. Shim rods, regulating rods, and safety rods are three different types of control rods.

Reactor Vessel

The reactor vessel is a tank-like structure that holds the reactor core and other internal components. The walls of the vessel are designed for a high-pressure radiation environment. In most cases the vessel walls are lined with thick steel slabs to reduce the flow of radiation from the core. As indicated in the last section, nuclear fission generates large amounts of neutrons and gamma rays, both of which are very harmful. Because of these, biological shielding is required around the reactor vessel. This shield consists of concrete blocks, which may be up to 2 m thick.

16.31.2 Types of Reactors

There are four types of reactors:

1. Pressurized water reactors (PWR)
2. Boiling water reactors (BWR)
3. Pressurized heavy-water reactor (PHWR)
4. Gas-cooled reactors (Magnox)

When ^{235}U is bombarded with neutrons, fission reactions take place, releasing neutrons that fission more atoms of ^{235}U. In order that the freshly released neutrons are able to fission further uranium atoms, their speeds must be reduced to a critical value. Therefore, for the reaction to be sustained, nuclear fuel rods must be embedded in neutron speed-reducing agents (such as graphite and heavy water) called moderators. For reaction control, rods made of neutron-absorbing material (boron steel) are used which, when inserted into the reactor vessel, control the amount of neutron flux, thereby controlling the range. A schematic diagram of a nuclear power plant is shown in Fig. 16.52. The heat released by the nuclear reaction is transported to a heat exchanger via primary coolant (CO_2, water etc.). Steam is then generated in the heat exchanger, which is used in a conventional manner to generate electric energy by means of a steam turbine. Various types of reactors are used in practice for power plant purposes, viz. advanced gas reactors (AGR), boiling water reactors (BWR), heavy-water-moderated reactors, etc.

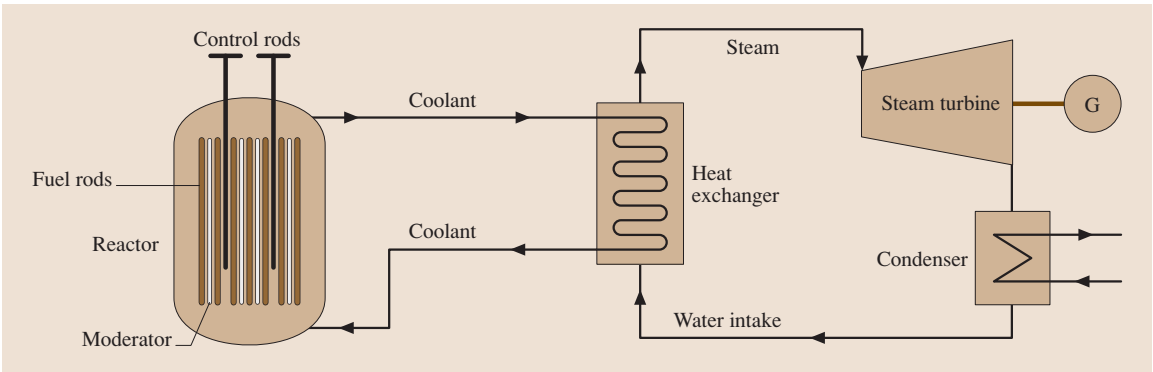


Fig. 16.52 Schematic view of a nuclear power plant

The CANDU reactor using natural uranium (in oxide from), moderated using pressurized heavy water is adopted in India. Its schematic diagram is shown in Fig. 16.53.

The associated merits and problems of nuclear power plants as compared to conventional thermal plants are discussed in the next paragraphs.

Merits and Demerits of Nuclear Power Plants

Merits. A nuclear power plant is totally free of air pollution and requires little fuel in terms of volume and weight; it therefore poses no transportation problems and may be sited, independently of nuclear fuel supplies, close to load centers. However, safety considerations require that they are normally located far away from populated areas.

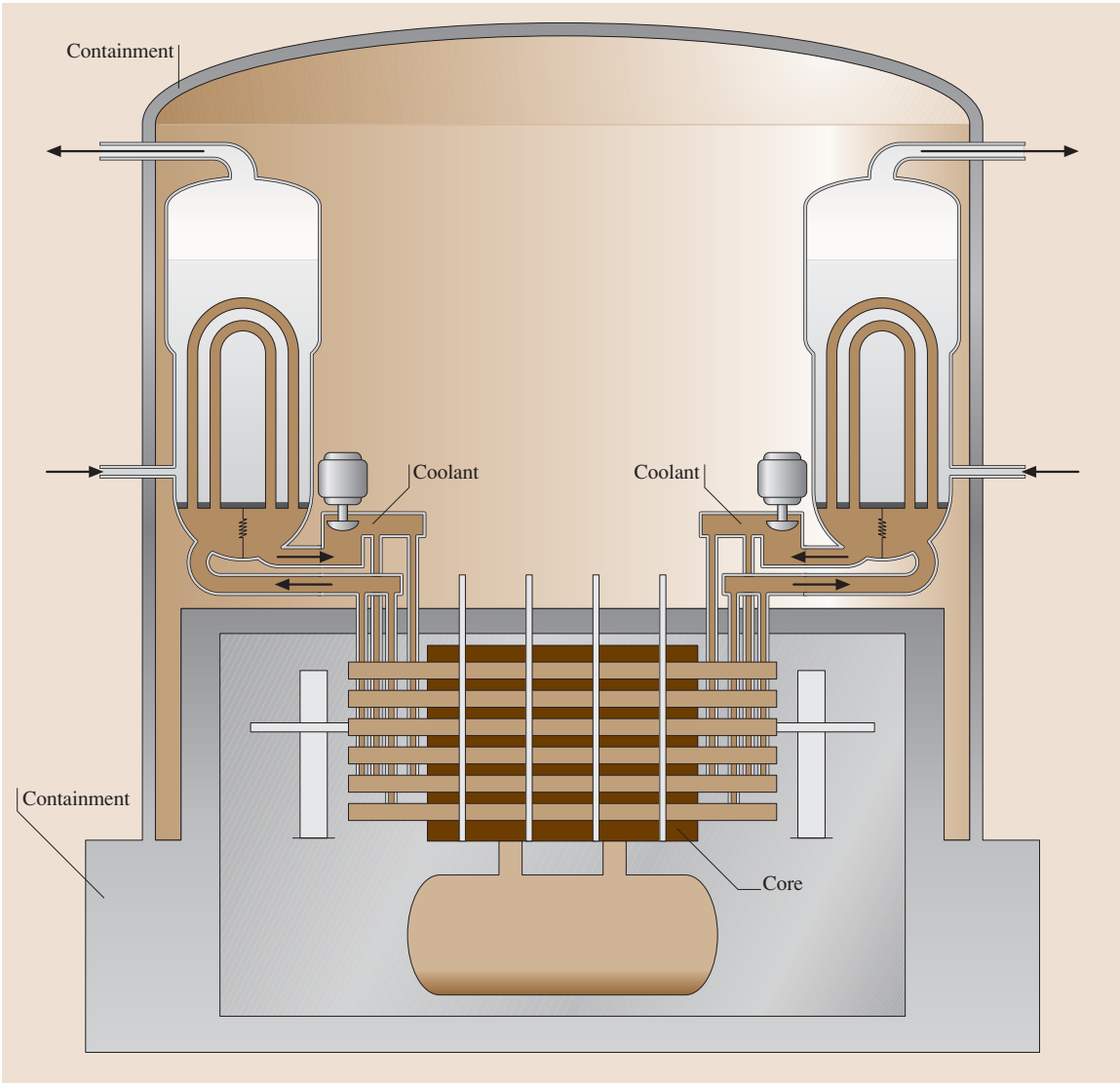


Fig. 16.53 CANDU reactor. Pressurized heavy-water-moderated design adopted in India

Demerits. Nuclear reactors produce radioactive fuel waste, the disposal of which poses serious environmental hazards. The rate of nuclear reaction can be lowered only by a small margin, so that the load on a nuclear power plant can only be marginally reduced below its full load value. Because of the relatively high capital cost against running cost, the nuclear plant should operate continuously as the base load station. Whenever possible, it is preferable to support such a station with a pumped storage scheme mentioned earlier. The greatest danger in a fission reactor is in the case of the loss of coolant in a accident. Even with the control rods fully lowered quickly, called a scram operation, fission continues and its afterheat may cause the vaporization and dispersal of radioactive material.

World uranium resources are quite limited, and at the present rate may not last much beyond 50 years. However, there is a redeeming feature. During the fission of ^{235}U some of the neutrons are absorbed by the more abundant uranium isotope ^{238}U (enriched uranium contains only about 3% of ^{235}U while most of it is ^{238}U) converting it to plutonium (^{239}U), which in itself

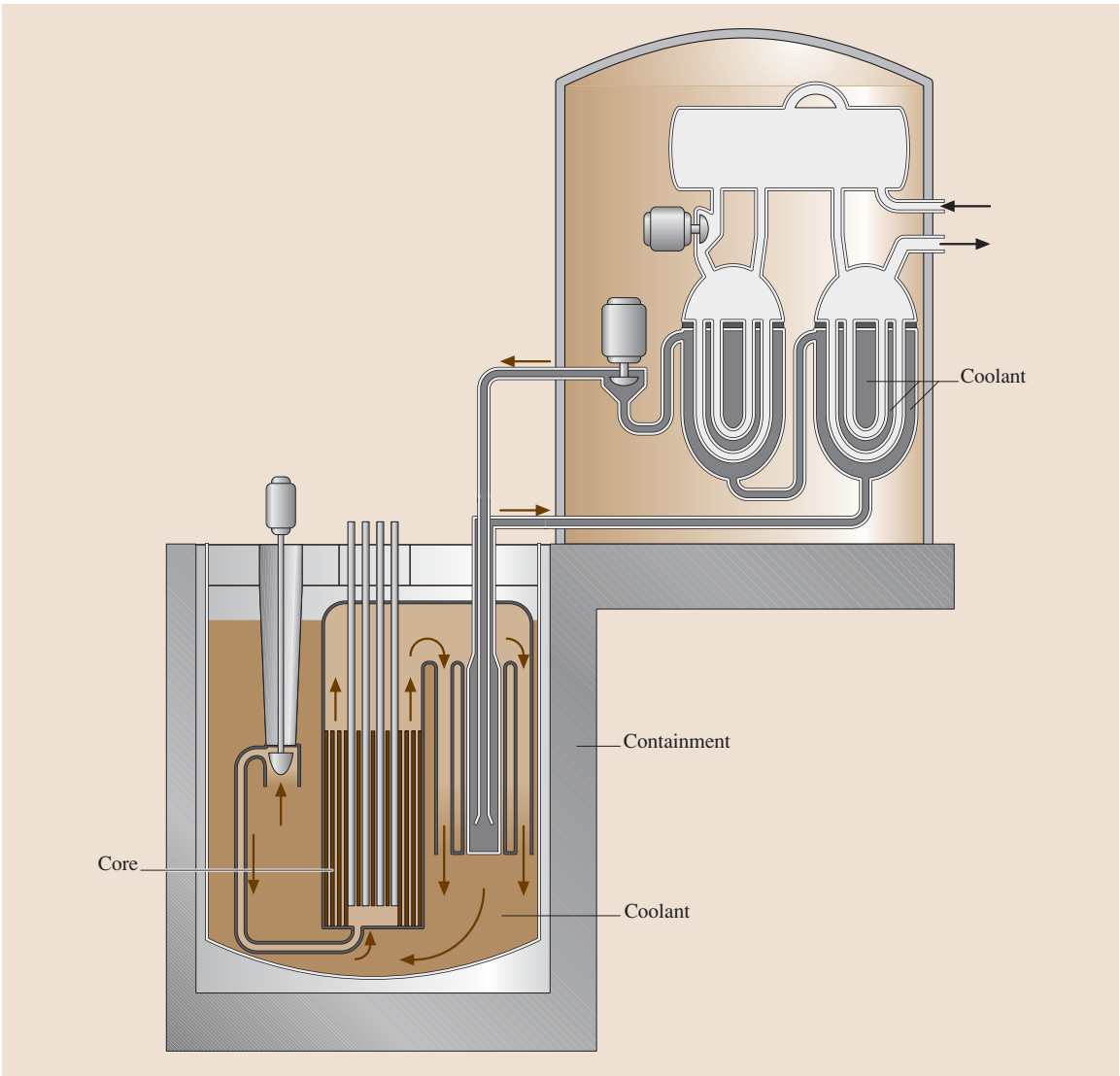


Fig. 16.54 A fast breeder reactor

is a fissionable material and can be extracted from the reactor fuel waste by a fuel reprocessing plant. Plutonium can then be used in the next-generation reactors (fast breeder reactors (FBRs)), thereby considerably extending the life of nuclear fuels. The FBR technology is being intensely developed as it will extend the availability of nuclear fuels at predicted rates of energy consumption to several centuries.

Figure 16.54 shows the schematic diagram of an FBR. It is essential that, for breeding operation, the conversion ratio (fissile material generated/fissile material consumed) is more than unity. This is achieved by fast-moving neutrons so that no moderator is needed, although the neutrons do slow down slightly through collision with structural and fuel elements. The energy density/kg of fuel is very high and so the core is small. It is therefore necessary that the coolant should possess good thermal properties; hence liquid sodium is used. The fuel for an FBR consists of 20% plutonium plus 8% uranium oxide. The coolant, liquid sodium, leaves the reactor at 650 °C at atmospheric pressure. The heat transported in this way is led to a secondary sodium circuit which transfers it to a heat exchanger to generate steam at 540 °C.

With a breeder reactor the release of plutonium, an extremely toxic material, make the environmental considerations very stringent.

An experimental fast breeder test reactor (FBTR) (40 MW) has been built at Kalpakkam alongside a nuclear power plant. FBR technology is expected to reduce the cost of electric energy so that it compares favorably with that from conventional thermal plants.

16.32 Future Prospects and Conclusion

Various advanced power generation technologies have been described in this chapter. Many regions of the world are experiencing fast growing electricity demand. Advanced technologies such as IGCC, ultra-supercritical cycles, and advanced gasification molten carbonate fuel cell cycles allow this electricity demand to be met and emission levels from power plants to meet air quality standards. Fluidized-bed combustion is an environmentally benign and proven technology for the disposal of solid wastes and the generation of electrical energy. Technological advancement will improve the reliability and efficiency of energy conversion process. The integrated gasification combined cycle will be able to exploit various kinds of low-grade energy resources such as biomass, low-grade coal, oil residues etc., for the sake of efficient power production. The development of advanced materials to withstand high temperatures as well as high pressures will enhance the thermal efficiency to 55%. Even a fraction of a percentage improvement in efficiency can mean huge savings in annual fuel cost.

Cogeneration of heat as well as electric power is one of the attractive options from the cost–benefit point of view; it saves 30–40% of fuel input energy. A combined cycle employing cogeneration using a multicomponent fluid such as an ammonia–water mixture will improve the thermal efficiency further. Effective utilization of renewable sources of energy including hydropower, solar, wind, and biomass is one of current the challenging tasks for researchers. The potential of solar power is unlimited; it is our prime task to utilize this energy in a more effective way. The development of solar technology seeks to achieve efficient operation even though solar energy intensity varies according to weather and time of day.

With the end of coal reserves in sight in the not too distant future, the immediate practical alternative source of large-scale electric energy generation is nuclear energy. The latest power plant technologies and dedicated research will lead to efficiencies approaching the Carnot efficiency in the near future.

References

16.1 A. Bejan, G. Tsatsaronis, M. Moran: *Thermal Design and Optimization* (Wiley, New York 1995)

16.2 D.P. Kothari, K.C. Singal, R. Ranjan: *Renewable Energy Sources and Emerging Technologies* (Prentice Hall of India, New Delhi 2008)

16.3 I. Seikan: *Steam Power Engineering: Thermal And Hydraulic Design Principles* (Cambridge Univ. Press, Cambridge 1999)

16.4 M.S. Briesch, R.L. Bannister, I.S. Diakunchak, A. Huber: A combined cycle designed to achieve

greater than 60% efficiency, ASME J. Eng. Gas Turbines Power **117**, 734–741 (1995)

16.5 R. Kehlhofer, R. Bachmann, H. Nielsen, J. Warner: *Combined Cycle Gas and Steam Turbine Power Plant* (PennWell, Tulsa 1999)

16.6 K.W. Li, A.P. Priddy: *Power Plant System Design* (Wiley, New York 1985)

16.7 D.P.K. Kothari: *Modern Power System Analysis* (McGraw-Hill, New York 2006)

16.8 R.C. Bansal, D.P. Kothari, T.S. Bhatti: On some of the design aspects of wind energy conversion systems, Int. J. Energ. Conv. Manag. **43**, 2175–2187 (2002)

16.9 D.P.K. Kothari, I.J. Nagarth: *Power System Engineering*, 2nd edn. (Tata McGraw-Hill, New Delhi 2007)

16.10 B. Kelly: Optimization studies for integrated solar combined cycle system, Proc. Solar Forum 2001 (ASME, Washington 2001)

16.11 J.C. Zink: Who says you can't store electricity, Power Eng. **101**(3), 21–25 (1997)

16.12 P.G. Hill: *Power Generation: Resources, Hazards Technology, and Costs* (MIT Press, Cambridge 1977)

16.13 W.A. Adams: Electrochemical energy storage systems: A small scale application to isolated communities in the Canadian Arctic, Can. Electr. Eng. J. **4**, 4–10 (1979)

16.14 S. Tavoulareas: *Multi Pollutant Emission Control Technology Options for Coal Fired Power Plants*, US Environmental Production Agency Rep. (EPA, Washington 2005)

16.15 L.E.J. Roberts, P.S. Liss, P.A.H. Saunders: *Power Generation and the Environment* (Oxford Univ. Press, Oxford 1990)

16.16 P. Basu, C. Kefa, L. Jestin: *Boilers and Burners: Design and Theory* (Springer, New York 2000)

16.17 G. Wills: *Nuclear Power Plant Technology* (Wiley, New York 1967)

17. Electrical Engineering

Seddik Bacha, Jaime De La Ree, Chris Oliver Heyde, Andreas Lindemann, Antje G. Orths, Zbigniew A. Styczynski, Jacek G. Wankowicz

Electricity is the most flexible form of energy accessible to humans. It can be transported over long distances, and transformed into almost any other kind of energy like heat, radiation or kinetic energy. Electrical engineering is very closely coupled especially to mechanical engineering but also to many other fields of engineering.

This chapter will give an overview of the theoretical fundamentals of electric phenomenon and some practical electric processes and application. It should be understood as a basic source of information about the most important issues in electrical engineering. For further reading the references will give a deeper insight into the mentioned scientific fields. The reader will get information about the fundamentals of electrical engineering in Sect. 17.1. Here the physical phenomenon of electric currents and voltages are explained. The electrical aspects of the main electrical machines transformer, generators and motors are explained in the Sects. 17.2, 17.3 and 17.5.

Power electronics have become a very important issue in transformation of different forms of electricity and in control of machinery. The reader will be informed about the basic working principles of this scientific field in Sect. 17.4.

This chapter places a emphasis on the section *Electric Power Transmission and Distribution* (17.6). In this section the fundamentals of electricity transport, distributed generation (especially from renewable sources) and the energy system protection are given.

17.1 Fundamentals

In electric engineering most of the phenomenon can be explained using some fundamental laws. These law are explained in Sect. 17.1.1. In the following subsections the basics about the electric circuits, starting from the DC current, and ending with the explanation of AC circuits with capacitance and inductance [17.1, 2]. The section closes with comments about multi phase networks and the behavior of some materials important for electrical engineering [17.1, 3].

17.1.1 Electric Field Basics

Fields and Equations

The electromagnetic field (EMF) in any given area of space should comply with the laws of electrodynamics [17.4]. This field can be described by the five field parameters presented in Table 17.1.

Based on experimental knowledge, four laws are known to be valid for electromagnetic fields: Maxwell's laws; in integral form they can be written as Ampère's, Faraday's, and Gauss's laws.

Ampère's Law. The first of Maxwell's laws is often called the general Ampère's law (17.1a)

$$\oint_C \boldsymbol{H}\,\mathrm{d}\boldsymbol{s} = \int\int_A \left(\boldsymbol{J} + \frac{\partial \boldsymbol{D}}{\partial t}\right)\mathrm{d}\boldsymbol{A} \tag{17.1a}$$

to which is added

$$I = \int\int_A \boldsymbol{J}\,\mathrm{d}\boldsymbol{A} \tag{17.1b}$$

and

$$\oint_C \boldsymbol{H}\,\mathrm{d}\boldsymbol{s} = I + \frac{\partial}{\partial t}\int\int_A \boldsymbol{D}\,\mathrm{d}\boldsymbol{A}\,. \tag{17.1c}$$

Ampère's law (17.1a) requires that the line integral (circulation) of the magnetic field intensity $\boldsymbol{H}$ around a closed contour $\boldsymbol{s}$ is equal to the sum of the line current I and electric displacement $\boldsymbol{D}$ through the surface $\boldsymbol{A}$.

Induction – Faraday's Law.

$$\oint_c \boldsymbol{E}\,\mathrm{d}\boldsymbol{s} = -\frac{\partial}{\partial t}\int\int_A \boldsymbol{B}\,\mathrm{d}\boldsymbol{A} \tag{17.2}$$

Faraday's law of induction requires that the line integral (circulation) of the electric field intensity $\boldsymbol{E}$ is equal to the negative time derivative of the magnetic flux density through the surface $\boldsymbol{A}$.

Magnetism – Gauss's Law.

$$\int\oint_A \boldsymbol{B}\,\mathrm{d}\boldsymbol{A} = 0 \tag{17.3}$$

Gauss's law states that the net magnetic flux out of any closed surface $\boldsymbol{A}$ is zero. This amounts to a statement about the sources of a magnetic field. For a magnetic dipole, in any closed surface the magnetic flux directed inward toward the south pole is equal to the flux outward from the north pole. The net flux will always be zero for dipole sources. If there were a magnetic monopole source, this would give a nonzero area integral. Furthermore, the divergence of a vector field is proportional to the point-source density, so the form of Gauss's law for magnetic fields is therefore a statement that there are no magnetic monopoles.

Electricity – Gauss's Law.

$$\int\oint_A \boldsymbol{D}\,\mathrm{d}\boldsymbol{A} = Q \tag{17.4}$$

The electric flux outside any closed surface is proportional to the charge Q.

The field parameters are related to three equations (17.5)–(17.7), which refer to materials of different physical properties

$$\boldsymbol{J} = \sigma\boldsymbol{E}\,, \tag{17.5}$$

$$\boldsymbol{D} = \varepsilon\boldsymbol{E}\,, \tag{17.6}$$

$$\boldsymbol{B} = \mu\boldsymbol{H}\,. \tag{17.7}$$

The current density and electric displacement are proportional to the electric field strength, while the flux

Table 17.1 Field parameters and symbols

	Field parameters	Symbol
Electric field	Electric displacement	$\boldsymbol{D}$
	Electrical field strength	$\boldsymbol{E}$
Magnetic field	Flux density	$\boldsymbol{B}$
	Magnetic field strength	$\boldsymbol{H}$
Electric flow field	Current density	$\boldsymbol{J}$

density is proportional to the magnetic field strength. The *electrical conductivity* σ, the *dielectric coefficient* ε and the *permeability* μ are generally tensors, but in isotropic substances may also be scalars.

The field equations are only applicable to rapidly changing processes, but can be adapted to slow processes, for application at typical technical frequencies. Each of these equations can be reduced for constant field parameters.

Electrostatic Field

Electrostatic field is a field of constant – in time – field parameters created by stationary charges. It is often called the electric field. Assuming constant field parameters and a static charge (on a dielectric surface), the following equation is valid

$$\oint_S \boldsymbol{E}\,\mathrm{d}\boldsymbol{s} = 0 \tag{17.8}$$

Equations (17.4), (17.6) and (17.8) together represent the basic relations for electrostatics.

The dielectric coefficient ε can be described as the product of the *permittivity of free space* ε_0 (vacuum) and the *relative permittivity* ε_r

$$\varepsilon = \varepsilon_0 \varepsilon_r \,, \tag{17.9}$$

where

$$\varepsilon_0 = 8.85 \times 10^{-12}\,\mathrm{A\,s/V\,m}\,.$$

The same is also valid for the permeability μ

$$\mu = \mu_0 \mu_\mathrm{r} \,, \tag{17.10}$$

with

$$\mu_0 = \frac{4\pi}{10} \times 10^{-6} = 1.25610^{-6}\,\mathrm{V\,s/A\,m}\,.$$

The electrical field strength (17.8) can also be described by (17.11).

$$\boldsymbol{E} = -\mathbf{grad}\,\varphi \,, \tag{17.11}$$

where ϕ is the scalar potential function.

From this, it is clear that the voltage is independent of the path of integration taken between two points and only depends on the difference of the potentials

$$\boldsymbol{V}_{12} = \int_1^2 \boldsymbol{E}\,\mathrm{d}\boldsymbol{s} = V_1 - V_2\,. \tag{17.12}$$

The electrostatic field can be illustrated using the equipotential lines, $\phi = \mathrm{const}$, and orthogonal field lines, which run over the tangential vector of the electrical field; see Fig. 17.1.

The electric force in an electrical field is defined as

$$\boldsymbol{F} = \int_Q \boldsymbol{E}\,\mathrm{d}Q \tag{17.13}$$

and can be created with a single point charge in the simplest case.

Consider two particles separated by a distance of $\boldsymbol{r}$ (in m) carrying charges of Q_1 and Q_2 coulombs, respectively. The force on Q_2 is given by

$$\boldsymbol{F}_2 = Q_2 \boldsymbol{E}\,, \tag{17.14}$$

where $\boldsymbol{E}$ is the electric field due to all charges except Q_2. In the present case there is only one charge other than Q_2. Let Q_1 be located at the origin of the spherical coordinates, then

$$\boldsymbol{E} = \frac{Q}{4\pi\varepsilon r^2}\boldsymbol{r}_0\,. \tag{17.15}$$

Using (17.15) for the electric field due to a point charge, we can deduce Coulomb's law

$$\boldsymbol{F}_2 = \boldsymbol{r}_0 \frac{Q_1 Q_2}{4\pi\varepsilon r^2}\,. \tag{17.16}$$

As one can see, the electric force between two charged particles is proportional to the product of the two charges and inversely proportional to the square of the distance between them. The unit vector $\boldsymbol{r}_0$ points into the direction of the connection line between two particles. If Q_1 and Q_2 are opposite in sign, the force is attractive, otherwise it is repulsive.

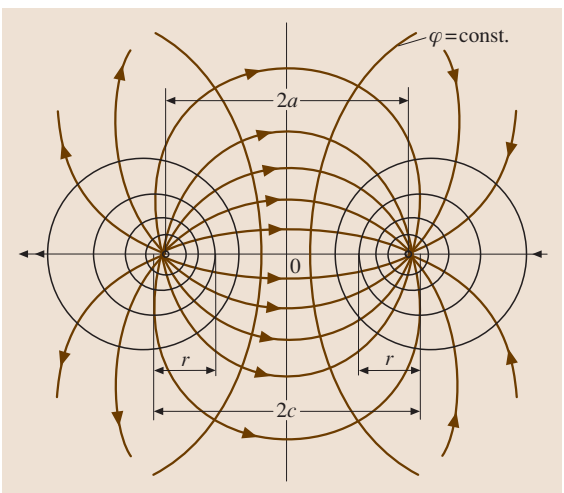


Fig. 17.1 Field characteristic [17.2]

Stationary Flow Field

Stationary flow field is an electric field in conducting surrounding. I an static stationary field all floating currents are constant in time. The current that passes through a surface is determined by the integral of the current density (17.17). The relationship between the electrical field strength $\boldsymbol{E}$ and current density $\boldsymbol{S}$ is given by (17.5).

$$I = \int\int_A \boldsymbol{S}\mathrm{d}\boldsymbol{A} \tag{17.17}$$

If a voltage V is present between two points of an electric circuit, the resistance R exists due to the constant conductivity σ. Based on (17.12) and (17.1b) the relation (17.18) known as Ohm's law can be written

$$R = \frac{V}{I} = \frac{1}{\sigma}\int_1^2 \boldsymbol{E}\,\mathrm{d}s \Big/ \int\int_A \boldsymbol{S}d\boldsymbol{A}\,. \tag{17.18}$$

The current flow generates losses in resistance, which turns into heat. Using *Joule's law* (Ohmic loss per volume p) this value can be determined exactly using

$$p = \sigma E^2 = (1/\sigma)S^2\,. \tag{17.19}$$

Stationary Magnetic Field

Stationary magnetic field is a field constant in time created by permanent magnets or direct current. Using Ampère's law for stationery conditions the path of integration of $\boldsymbol{H}$ along the contour $\boldsymbol{s}$ is equal to the total current I that flows through this area

$$\oint_S \boldsymbol{H}\,\mathrm{d}\boldsymbol{s} = I\,. \tag{17.20}$$

The relationship between field strength and flux density is still valid

$$\boldsymbol{B} = \mu\boldsymbol{H}\,.$$

Lastly, the characteristic of magnetization (flux density $\boldsymbol{B}$ across field strength) is nonlinear for ferromagnetic materials.

17.1.2 Electric Circuits

Direct-Current (DC) Circuits

This section discusses circuits composed of direct-voltage sources and ohmic resistance, which are connected by electrical conductors. The resistance through which the electric current flows causes a voltage drop, which is defined using (17.18)

$$V = RI$$

and, respectively

$$I = GV\,, \tag{17.21}$$

where I is the direct current, that is characterized by constant in time value and sense while $G = 1/R$ is the reciprocal of the resistance R, usually called the conductance.

The electric power P is the product of current I and voltage V and is defined as

$$P = VI = RI^2 = V^2/R\,. \tag{17.22}$$

The direct-voltage source can be presented as an ideal source (ideal voltage V source) connected in series with an internal resistance R_s, e.g., a battery. The source can also be represented as a load-independent DC current source I_s connected in parallel to an internal resistance R_s, see Fig. 17.2.

Under the condition that current density $\boldsymbol{S}$ is a constant value the resistance R can be also calculated on the basis of parameters that describe geometrical and material properties of a conductor such as length l, cross-section A, and resistivity ρ

$$R = \rho l/A\,. \tag{17.23}$$

The parameter $\rho = 1/\kappa$ is called the material resistivity and depends generally on physical properties

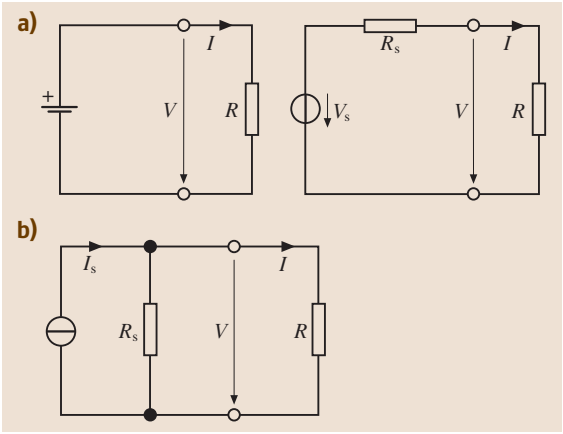


Fig. 17.2a,b Voltage and current sources: (a) imposed voltage; (b) imposed current

of the material and its temperature. It has a linear characteristic for many materials. The resistivity ρ_{20} – determined at a temperature of 20 °C – is a reference value for calculations of the parameter ρ using (17.24). The resistivity ρ_{20} is a material constant.

$$\rho = \rho_{20}(1 + \alpha(\vartheta - 20\,^{\circ}\mathrm{C}))\,, \tag{17.24}$$

where ϑ is the actual temperature of a conductor and α is the specific temperature coefficient.

Kirchhoff's Laws

- Kirchhoff's current law (KCL). The first Kirchhoff law says that, at each node in any electrical network and at each moment of time, the algebraic sum of all currents leaving a node is zero

$$\sum_{i=1}^{n} I_i = 0\,. \tag{17.25}$$

- Kirchhoff's voltage law (KVL). One of the most significant aspects of series circuits is described by the second Kirchhoff law, which states that the algebraic sum of the voltage rises and drops over all elements in any closed loop is equal to zero

$$\sum_{i=1}^{n} V_i = 0 \tag{17.26}$$

Using these two equations the values of parallel and serial connected resistances can be calculated:

- For serial connection

$$R_{\mathrm{res}} = \sum_{i=1}^{n} R_i \tag{17.27}$$

- For parallel connection

$$G_{\mathrm{res}} = \sum_{i=1}^{n} G_i = \sum_{i=1}^{n} \frac{1}{R_i} = \frac{1}{R_{\mathrm{res}}} \tag{17.28}$$

Kirchhoff's laws can also be used as variables in time-dependent currents and voltages. Taken together, these laws have created the basis of network theory.

Example 17.1: The calculation of the cross current in bridge circuits

Let us now consider a bridge circuit as an example circuit. In Fig. 17.3 the load-independent voltage source supplies the circuit, which is composed of the

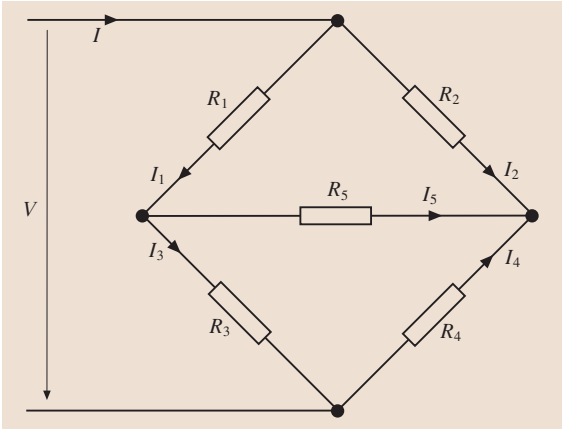


Fig. 17.3 Bridge circuit

bridge sections $R_1, \ldots, R_4$ and one cross branch R_5. The network has four nodes $n = 4$, and three independent equations $(n - 1 = 3)$ can be created on the basis of the first Kirchhoff law (17.24).

$$\begin{aligned} I &= I_1 + I_2\,, \\ I &= I_3 + I_4\,, \\ I_5 &= I_1 - I_3\,. \end{aligned} \tag{17.29}$$

On the basis of the second Kirchhoff law (17.26) a next set of equations can be written

$$\begin{aligned} 0 &= R_1 I_1 + R_3 I_3 - V\,, \\ 0 &= R_1 I_1 + R_5 I_5 - R_2 I_2\,, \\ 0 &= R_3 I_3 - R_5 I_5 + R_4 I_4\,. \end{aligned} \tag{17.30}$$

Taking into account these equations the current in the cross branch can be calculated as

$$\begin{aligned} I_5 = V(R_2 R_3 - R_1 R_4) \big/ [&R_5(R_1 + R_3)(R_2 + R_4) \\ &+ R_1 R_3(R_2 + R_4) + R_2 R_4(R_1 + R_3)]\,. \end{aligned} \tag{17.31}$$

The cross current I_5 disappears in a balanced bridge. This is valid when

$$\frac{R_1}{R_2} = \frac{R_3}{R_4}\,. \tag{17.32}$$

Example 17.2: Converting a star connection into a delta connection

This transformation is often used to simplify the calculation of larger networks.

In Fig. 17.4 the star connection with resistances R_1, R_2, and R_3 follows the same currents and voltage drops

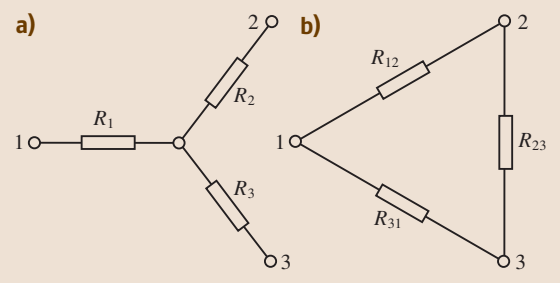


Fig. 17.4a,b Connection types: (a) star connection; (b) delta connection

with regard to points 1, 2, and 3 as the delta connection with resistances R_{12}, R_{23}, and R_{31} only when the following equations are fulfilled

$$R_1 = \frac{R_{31}R_{12}}{R_{12}+R_{23}+R_{31}}\,, \tag{17.33}$$

$$R_{12} = R_1 + R_2 + \frac{R_1 R_2}{R_3}\,, \tag{17.34}$$

$$R_2 = \frac{R_{12}R_{23}}{R_{12}+R_{23}+R_{31}}\,, \tag{17.35}$$

$$R_{23} = R_2 + R_3 + \frac{R_2 R_3}{R_1}\,, \tag{17.36}$$

$$R_3 = \frac{R_{23}R_{31}}{R_{12}+R_{23}+R_{31}}\,, \tag{17.37}$$

$$R_{31} = R_3 + R_1 + \frac{R_3 R_1}{R_2}\,. \tag{17.38}$$

Capacitor

The relationship between the capacitance C, charge Q, and voltage V is given

$$C = \frac{Q}{V} = \int\limits_A \int \boldsymbol{D}\,\mathrm{d}\boldsymbol{A} \Big/ \int\limits_D \boldsymbol{E}\,\mathrm{d}s\,. \tag{17.39}$$

The simplest example of a capacitor (Fig. 17.5) is the parallel plate capacitor. It consists of two conductive plate shaped electrodes separated by a dielectric.

By disregarding the side effects and using (17.5–17.7) we can define the following relationships

$$\varphi = \frac{Q}{\varepsilon A}x\,, \tag{17.40}$$

$$E = \frac{Q}{\varepsilon A} = \frac{V}{d}\,, \tag{17.41}$$

$$C = \frac{Q}{V} = \varepsilon\frac{A}{d}\,. \tag{17.42}$$

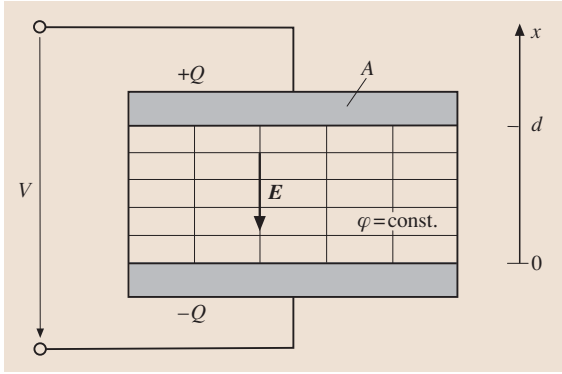


Fig. 17.5 Principle circuit of capacitor [17.2]

The equations (17.39–17.41) are also suitable for variable charge q and voltage v

$$i = \frac{\mathrm{d}q}{\mathrm{d}t} = C\frac{\mathrm{d}v}{\mathrm{d}t}\,. \tag{17.43}$$

The energy W_S in the capacitor can be generally calculated by the formula

$$W_s = \int\limits_0^V q\,\mathrm{d}v = \frac{1}{2}CV^2\,. \tag{17.44}$$

Faraday's Law

Faraday's law is a fundamental relationship which comes from Maxwell's equations. It serves as a succinct summary of the ways a voltage (or EMF) may be generated by a changing magnetic environment. The induced EMF in a coil is equal to the negative (Lenz's law) of the rate of change of magnetic flux times the number of turns in the coil. It involves the interaction of the charge with the magnetic field.

This law shows exactly that any change in the magnetic environment of a coil of wire will cause a voltage (electromagnetic field) to be *induced* in the coil. Regardless of which way the change is produced, a voltage will be generated. The change could be produced by changing the magnetic field strength, moving a magnet toward or away from the coil, moving the coil into or out of the magnetic field, rotating the coil relative to the magnet, etc.

$$v_i = -N\frac{\mathrm{d}\Phi}{\mathrm{d}t}\,, \tag{17.45}$$

where N is the number of turns and Φ is the magnetic flux.

Inductor

An inductor is a two-terminal device that consists of coiled conducting wire. A current flowing through the device produces a magnetic flux which forms closed loops encircling the coils making up the inductor. Suppose that the coil contains N turns and that the flux passes through each turn.

In a linear inductor, the magnetic flux linkage ψ is directly proportional to the current flowing through the device.

The equation (17.46) defines the so-called inductance L with the unit Henry

$$L = \frac{\psi}{I} = \frac{\sum \Phi_n}{I} \,. \tag{17.46}$$

Using (17.45) and (17.46), an important relationship for inductors can be obtained

$$V = L\frac{\mathrm{d}i}{\mathrm{d}t} \,. \tag{17.47}$$

Clearly, as i increases, a voltage is generated between the terminals of the inductor, the polarity of which is shown in Fig. 17.6. This voltage opposes an increase in i, as if this were not the case and the polarity was reversed, the induced voltage would *aid* the current.

Just as work was performed in moving charges between the plates of a capacitor, work is necessary to establish the flux in the inductor. The work energy required in this case is said to be stored in the magnetic field.

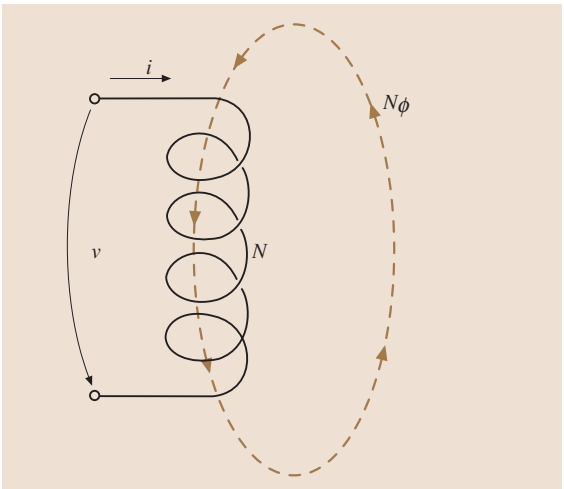


Fig. 17.6 Principle of an inductor

The energy stored in an inductor is given by

$$W_\mathrm{m} = \int_0^I \Phi \,\mathrm{d}i = \frac{1}{2} L I^2 \,. \tag{17.48}$$

Magnetic Materials

The behavior of materials in a magnetic field can be described as:

- Paramagnetic
- Diamagnetic
- Ferromagnetic

The permeability μ_r of the first two types of materials is approximately unity. However, ferromagnetic materials, particularly iron, nickel, and cobalt behave differently. In the presence of a magnetic field they produce con-

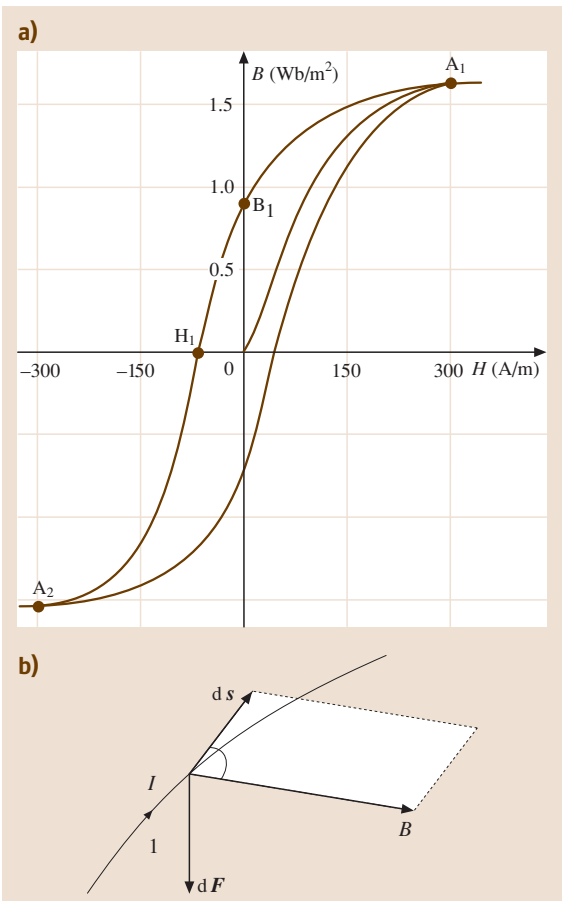


Fig. 17.7 **(a)** Hysteresis curve; **(b)** conductor l leading current I

siderably more magnetic flux density than air. The magnetic flux density B reinforcement is expressed by

$$B = J + \mu_0 H = \mu_0 (M + H) , \tag{17.49}$$

where J is the magnetic polarization and M is the magnetization.

Ferromagnetic materials show magnetization characteristics in the form of a hysteresis curve $B = f(H)$, see Fig. 17.7a. After initial magnetization to the point A_1, the H field is decreased by reducing the current in the coil, the magnetic flux density B is also reduced, but the B–H curve does not follow the original magnetization curve. Note that at point B_1 the current in the coil of a toroid is equal to zero, as is H, yet the residual magnetic flux density B is nonetheless present in the toroid.

The residual flux is due to the fact that the magnetic moments of some of the domains in the ferromagnetic materials are still aligned in the same direction. The magnitude of this residual B is called the remanence. If we now reverse the current to reverse the H, the B–H curve will trace a curve $B_1 H_1$, as shown in Fig. 17.7. Note that it takes some negative value of H to null the B field. The value of H_1 in the negative direction of the initial magnetization, necessary to nullify the magnetic inductance B field, is called the coercive force or coercivity. If the reverse current is increased beyond this point, the magnetic inductance B begins to reverse and the B–H curve follows the curve $H_1 A_2$. If the current is now reduced, the B–H curve traces a new path $A_2 A_1$. The closed loop $A_1 B_1 H_1 A_2 A_1$ is called a hysteresis loop. If the current is varied in a smaller cycle, the corresponding hysteresis loop will be smaller. If the material is brought to saturation in both ends of the magnetization curve, the remanence B is called the retentivity of the ferromagnetic material, and the coercive force H is called the coercivity of the material.

In general both soft and hard magnetic materials are useful in electrical engineering. The first are used for constructing magnetic circles in electrical machines (slim hysteresis curve), while hard magnetic materials are used in permanent magnets.

Electromagnetic Field Forces

An electromagnetic field can be defined by the force it produces. Generally, the force in a field is defined as

$$\boldsymbol{f}_v = (\boldsymbol{S} \times \boldsymbol{B}) - \frac{1}{2} \boldsymbol{H}^2 \mathbf{grad}\, \mu . \tag{17.50}$$

The first part of (17.50) gives the current density force, which affects the current leading conductor in the external field from B induction. The second part appears only for position-dependent permeability and is expressed as a permeable force density.

For the length element $\mathrm{d}\boldsymbol{s}$ of an electrical conductor the force is expressed as

$$\mathrm{d}\boldsymbol{F} = I(\mathrm{d}\boldsymbol{s} \times \boldsymbol{B}) . \tag{17.51}$$

The force is determined for a cable with length l, which is located in a vertical magnetic field (Fig. 17.7b)

$$F = BIl . \tag{17.52}$$

The so-called *right-hand rule* is applied to determine the direction of the magnetic force.

A magnetic force can also be specified for two parallel electrical conductors. The current I_1 generates the field strength H at a distance r in conductor 1, in agreement with (17.1a)

$$H = I_1/(2\pi r) . \tag{17.53}$$

A second conductor located at the distance d and carrying the current I_2 is affected by the force

$$F = \frac{\mu_0}{2\pi} \frac{I_1 I_2}{d} l . \tag{17.54}$$

The value of the forces in conductors 1 and 2 is the same. When both the current in the two conductors is in the same direction electrostatic attraction occurs. If, on the other hand, the direction of the current in the two conductors is different then repulsion occurs.

Moreover, mechanical tensions are produced on the interface of a magnetic field, between areas with differential permeability. On the interface between iron and air longitudinal pull and lateral pressure develop.

If a magnetic field with inductance B passes vertically through a surface that combines areas with μ_1 and μ_2, a specific force σ is accrued normal to this surface

$$\sigma = \frac{1}{2} \left(\frac{1}{\mu_1} - \frac{1}{\mu_2} \right) B^2 . \tag{17.55}$$

When $\mu_1 = \mu_0$ for air and $\mu_2 = \mu_0 \mu_r \gg \mu_0$ for iron the electrostatic attraction force through surface A can be defined by

$$F = \frac{1}{2\mu_0} B^2 A . \tag{17.56}$$

This is applied in electromagnets carrying ferromagnetic loads.

17.1.3 Alternating Current (AC) Engineering

Alternating Current Quantities

For a periodic current waveform $i(t)$ with period T and frequency $f = 1/T$ the following definitions can be written

- The average value

$$\bar{i} = \frac{1}{T}\int_0^T i\,\mathrm{d}t \tag{17.57}$$

- The effective, root-mean-square (RMS) value

$$I = \sqrt{\frac{1}{T}\int_0^T i^2\,\mathrm{d}t} \tag{17.58}$$

The average value $\bar{v}$ and the RMS value V of an alternating voltage $v(t)$ are defined in a similar way.

The AC waveform current containing a DC component consists of a fundamental frequency and higher harmonics that are integer multiples of the fundamental frequency. The RMS value of such a current is

$$I = \sqrt{\bar{i}^2 + I_1^2 + I_2^2 + I_3^2 + \ldots} = \sqrt{\bar{i}^2 + I_{\approx}^2}\,. \tag{17.59}$$

The deformation of an alternating current through harmonics can be calculated using the distortion factor

$$k_\mathrm{i} = \sqrt{1 - \left(\frac{I_\mathrm{i}}{I}\right)^2}\,. \tag{17.60}$$

In this case a fundamental current will be defined as

$$i = \hat{i}\cos(\omega t + \varphi)\,, \tag{17.61}$$

where $\hat{i}$ is the peak value of the current and $\omega = 2\pi f$ is the angular speed.

The average value $\bar{i}$ of the current i (17.57) is zero while the effective value is

$$I = \hat{i}/\sqrt{2}\,. \tag{17.62}$$

The same refers to voltages and therefore

$$V = \frac{\hat{v}}{\sqrt{2}}\,. \tag{17.63}$$

RMS value of a current I refers directly to the Ohmic resistance by thermal energy dissipated by this both at flow of direct current or its alternating equivalent of defined RMS value which produces the same thermal effect.

Active power P produced by current I of given RMS value can be written as

$$P_\mathrm{V} = R\frac{1}{T}\int_0^T i^2\,\mathrm{d}t = RI^2\,. \tag{17.64}$$

Currents like that in (17.61) flow in branches with linear elements if the applied voltages and currents also have a sine waveform with a fundamental frequency, and transition processes have faded away.

Alternating Current Parameters

Every alternating current parameter is represented by its:

- Amplitude
- Effective value
- Frequency
- Phasing ($t = 0$)

Applying these parameters the current and voltage can be expressed as follows

$$i = \mathrm{Re}\left(\sqrt{2}I\mathrm{e}^{\mathrm{j}(\omega t+\varphi_i)}\right) = \mathrm{Re}\left(\sqrt{2}\underline{I}\mathrm{e}^{\mathrm{j}\omega t}\right)\,, \tag{17.65}$$

with $\underline{I} = I\mathrm{e}^{\mathrm{j}\varphi_\mathrm{i}}$

$$v = \mathrm{Re}\left(\sqrt{2}V\mathrm{e}^{\mathrm{j}(\omega t+\varphi_u)}\right) = \mathrm{Re}\left(\sqrt{2}\underline{V}\mathrm{e}^{\mathrm{j}\omega t}\right)\,, \tag{17.66}$$

with $\underline{V} = V\mathrm{e}^{\mathrm{j}\varphi_u}$.

Now, by giving the angular frequency ω the current and voltage can be sufficiently described by using the complex parameters $\underline{I}$ and $\underline{V}$, which contain information about the absolute value and phasing.

The ratio of $\underline{I}$ to $\underline{V}$ is complex and denotes the impedance $\underline{Z}$, which consists of the component's resistance R and reactance X. Its multiplicative inverse gives the admittance $\underline{Y}$, which consists of the conductance G and the susceptance B

$$\underline{Z} = \underline{V}/\underline{I} = R + \mathrm{j}X\,, \tag{17.67}$$

$$\underline{Y} = \underline{I}/\underline{V} = 1/\underline{Z} = G + \mathrm{j}B\,. \tag{17.68}$$

The passive linear elements in alternating current circuits are:

- Ohmic resistance
- Capacitors
- Inductors

By using (17.65)–(17.68) it is very easy to evaluate these elements for AC

$$v_R = Ri \rightarrow V_R = RI \rightarrow Z_R = R\,, \tag{17.69}$$

$$i_\mathrm{c} = C\frac{\mathrm{d}v}{\mathrm{d}t} \rightarrow I_\mathrm{c} = \mathrm{j}\omega CV \rightarrow Z_\mathrm{c}\,,$$

$$\underline{Z}_\mathrm{c} = 1/\mathrm{j}\omega C = -\mathrm{j}X_\mathrm{c}\,, \tag{17.70}$$

$$v_\mathrm{l} = L\frac{\mathrm{d}i}{\mathrm{d}t} \rightarrow V_\mathrm{L} = \mathrm{j}\omega LI \rightarrow Z_\mathrm{L}\,,$$

$$\underline{Z}_\mathrm{L} = \mathrm{j}\omega L = \mathrm{j}X_\mathrm{L}\,. \tag{17.71}$$

Accordingly, in relation to the voltage the current phase is shifted by $-\pi/2$ by an inductor, and by $+\pi/2$ by a capacitor. A resistance does not produce an angular phase shift.

Electric Power

In one-phase circuits the momentary value of electric power is expressed by

$$p(t) = v(t)i(t)\,. \tag{17.72}$$

When the current and voltage have a sine waveform one can use (17.65) and (17.66) to deduce

$$\begin{aligned} p(t) &= VI[\cos\varphi + \cos(2\omega t + \varphi_\varepsilon)] \\ &= P + S\cos(2\omega t + \varphi_\varepsilon)\,, \end{aligned} \tag{17.73}$$

where $\varphi = \varphi_u - \varphi_i$ and $\varphi_\varepsilon = \varphi_u + \varphi_i$.

P is the active power, Q the reactive power, and S the apparent power. Some additional important equations (Fig. 17.8b) are given below

$$P = VI\cos\varphi\,, \tag{17.74}$$

$$S = VI\,, \tag{17.75}$$

$$Q = \sqrt{S^2 - P^2} = VI\sin\varphi\,. \tag{17.76}$$

In complex form, the power can be written

$$S = P + \mathrm{j}Q\,. \tag{17.77}$$

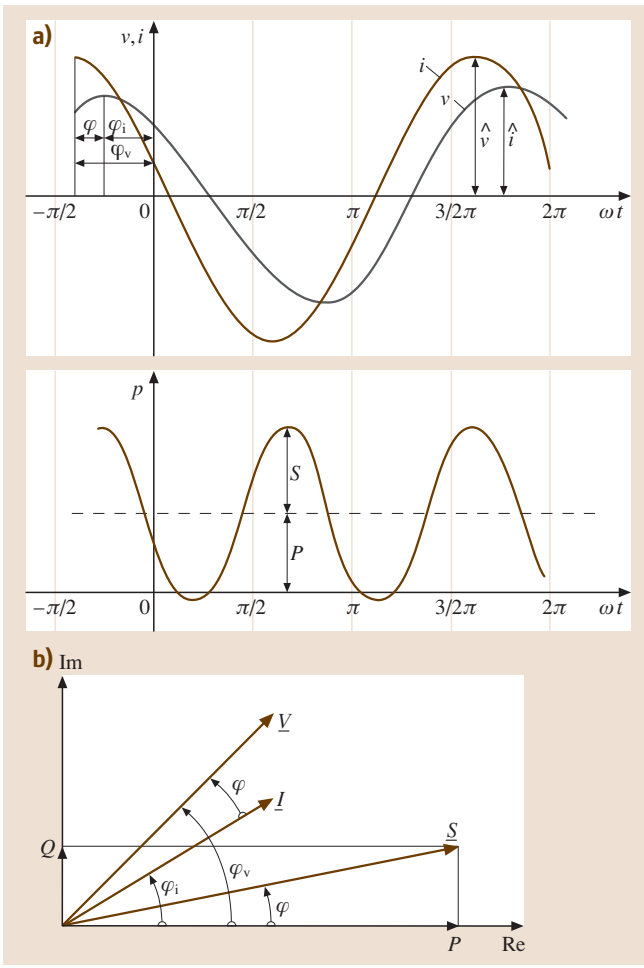


Fig. 17.8a,b Alternating values diagram: **(a)** characteristics of current, voltage, and power; **(b)** phasor diagram [17.2]

Three-Phase Current

The three-phase current system can be connected in one of two topologies: star or delta connection (Fig. 17.9).

The three-phase system is symmetrical when its values have equal amplitudes with equal frequencies and every phase is delayed by $2\pi/3$.

This can be applied to the voltage system

$$\begin{aligned} v_1 &= \sqrt{2}V\cos\omega t \rightarrow \underline{V}_1 = V\,, \\ v_2 &= \sqrt{2}V(\cos\omega t - 2\pi/3) \rightarrow \underline{V}_2 = V\,\mathrm{e}^{-\mathrm{j}2\pi/3}\,, \\ v_3 &= \sqrt{2}V(\cos\omega t - 4\pi/3) \rightarrow \underline{V}_3 = V\,\mathrm{e}^{\mathrm{j}2\pi/3}\,. \end{aligned} \tag{17.78}$$

Generally, the three phases are labeled as follows:

- U (1 phase), L_1 – line 1
- V (2 phase), L_2 – line 2
- W (3 phase), L_3 – line 3

In Fig. 17.9 the symmetrical star and delta connections are shown. The line-to-line voltages V_{12}, V_{23}, and V_{31}, and the line-to-line currents I_1, I_2, and I_3, together follow the relations given below.

Star Connection

The line-to-line currents are equal to the line-to-neutral currents

$$I_1 = I_{1N}\,; \quad I_2 = I_{2N}\,; \quad I_3 = I_{3N}\,.$$

The line-to-line voltages are equal to the differences of the line-to-neutral voltages

$$\begin{aligned} V_{12} &= V_{1N} - V_{2N} \\ V_{23} &= V_{2N} - V_{3N} \\ V_{31} &= V_{3N} - V_{1N}\,. \end{aligned} \tag{17.79}$$

If they are symmetrical

$$\begin{aligned} I_\mathrm{L} &= I_\mathrm{iN}\,, \\ V_\mathrm{L} &= \sqrt{3}V_\mathrm{iN}; \end{aligned}$$

as well as

$$I_N = -I_1 - I_2 - I_3 = 0 \,.$$

The neutral line does not carry a current and is dispensable.

Delta Connection

The line-to-line currents are equal to the difference between the line-to-neutral currents

$$\begin{aligned} I_1 &= I_{12} - I_{31} \\ I_2 &= I_{23} - I_{12} \\ I_3 &= I_{31} - I_{23} \,. \end{aligned} \tag{17.80}$$

The line-to-line voltages are equal to the line-to-neutral voltages

$$\begin{aligned} V_{12} &= V_{UV} \\ V_{23} &= V_{VW} \\ V_{31} &= V_{WU} \,. \end{aligned} \tag{17.81}$$

If they are symmetrical

$$\begin{aligned} V_L &= V_{iN} \,, \\ I_L &= \sqrt{3} I_{iN} \,. \end{aligned} \tag{17.82}$$

In three-phase systems the power is independent of the circuit type

$$P = \sqrt{3} V_L I_L \cos\varphi = S \cos\varphi \,. \tag{17.83}$$

The active power is constant in time, and the reactive power is defined as

$$Q = \sqrt{3} V_L I_L \sin\varphi = S \sin\varphi \,. \tag{17.84}$$

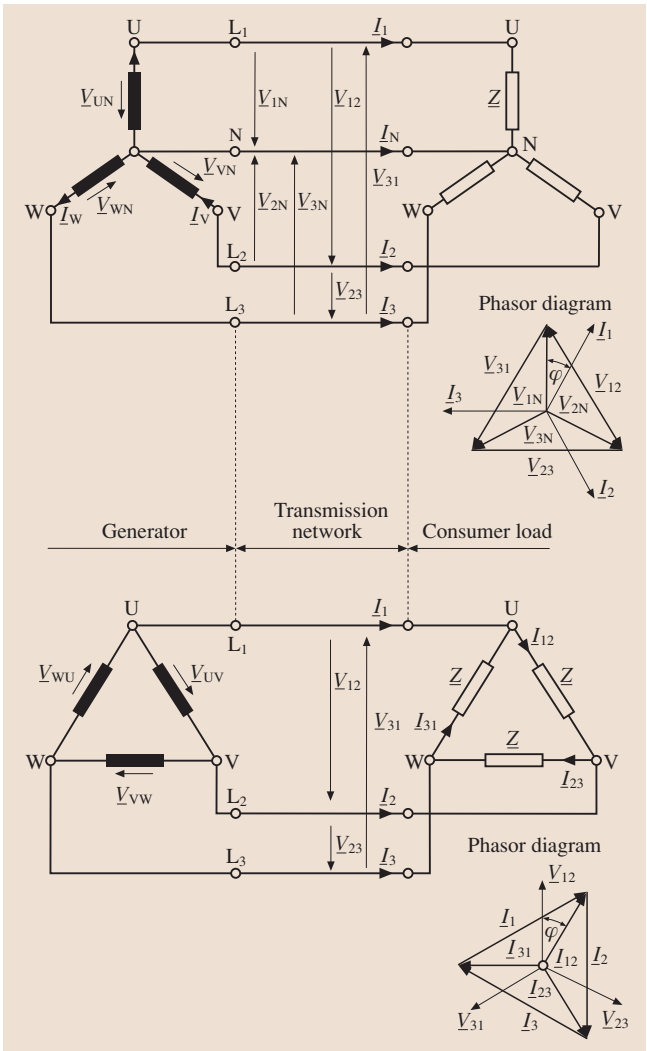


Fig. 17.9 Symmetrical three-phase system in the star and delta connections

Symmetrical Components. Figure 17.10 shows the systems with symmetrical currents and voltage, which are characterized by equal amplitudes and a delay of each phase by $2\pi/3$. An asymmetrical load, and especially asymmetrical short circuits, create an asymmetrical system. The method of symmetrical components will be used to analyze the circuits in this section.

Symmetrical components are used to analyze non symmetrical network conditions. Usually networks are supposed to work symmetrical, which means that loads are distributed equally to all phases and no disturbance takes place. However, if an unsymmetrical situation occurs, the method of symmetrical components is applied. This method allows for separation of the unsymmetrical phasors into three systems of symmetrical phasors.

The equations (17.85) and (17.86) show the transformation to symmetrical components and the back transformation. The three systems can be described as follows:

- Positive sequence system: three phasors each shifted by 120°, rotating in the same direction as the original phasors (index 1)

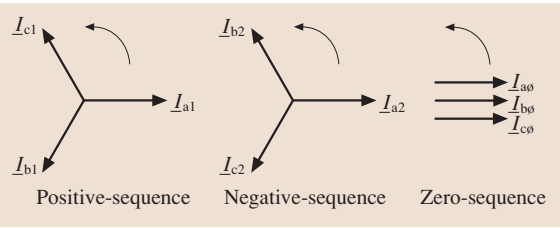


Fig. 17.10 Symmetrical components

- Negative sequence system: three phasors each shifted by 120°, rotating in the same direction but with different phase sequence (acb) (index 2)
- Zero sequence system: all phasors point to the same direction (index 0)

$$\begin{pmatrix} \underline{I}_0 \\ \underline{I}_1 \\ \underline{I}_2 \end{pmatrix} = \frac{1}{3} \begin{pmatrix} 1 & 1 & 1 \\ 1 & \underline{a} & \underline{a}^2 \\ 1 & \underline{a}^2 & \underline{a} \end{pmatrix} \begin{pmatrix} \underline{I}_\mathrm{U} \\ \underline{I}_\mathrm{V} \\ \underline{I}_\mathrm{W} \end{pmatrix} , \qquad (17.85)$$

where $a = \mathrm{e}^{\mathrm{j}2\pi/3}$

$$\begin{pmatrix} \underline{I}_\mathrm{U} \\ \underline{I}_\mathrm{V} \\ \underline{I}_\mathrm{W} \end{pmatrix} = \frac{1}{3} \begin{pmatrix} 1 & 1 & 1 \\ 1 & \underline{a}^2 & \underline{a} \\ 1 & \underline{a} & \underline{a}^2 \end{pmatrix} \begin{pmatrix} \underline{I}_0 \\ \underline{I}_1 \\ \underline{I}_2 \end{pmatrix} . \qquad (17.86)$$

Furthermore, phase currents from symmetrical components can be recalculated each time.

Now an unsymmetrical system is composed of three symmetrical systems which can be analyzed with the methods explained above. Unsymmetrical system conditions are mainly caused by unsymmetrical short-circuits or heavily unsymmetrical loads.

The symmetrical components method is effective in short-circuit calculations of electrical machines and networks.

Transfer Locus Diagram. A transfer locus for one complex surface is a geometrical area for the end points of all phasors and depends on one real parameter. which is usually the angular frequency ω, but may also be the frequency of oscillation.

Taking into account the circuit in Fig. 17.11a, the impedance transfer locus will be represented as $\underline{Z} = R + \mathrm{j}\omega L$.

The result for the admittance $\underline{Y} = 1/Z$ is a semicircle, because $\underline{Z}(\omega)$ is only valid for positive imaginary values.

By suitable scale fitting and fixed applied voltage $\underline{V}$, the $\underline{Y}(\omega)$ locus also represent the current locus $\underline{I}(\omega)$. In this case the voltage drops on resistance and inductance add to the supplied voltage.

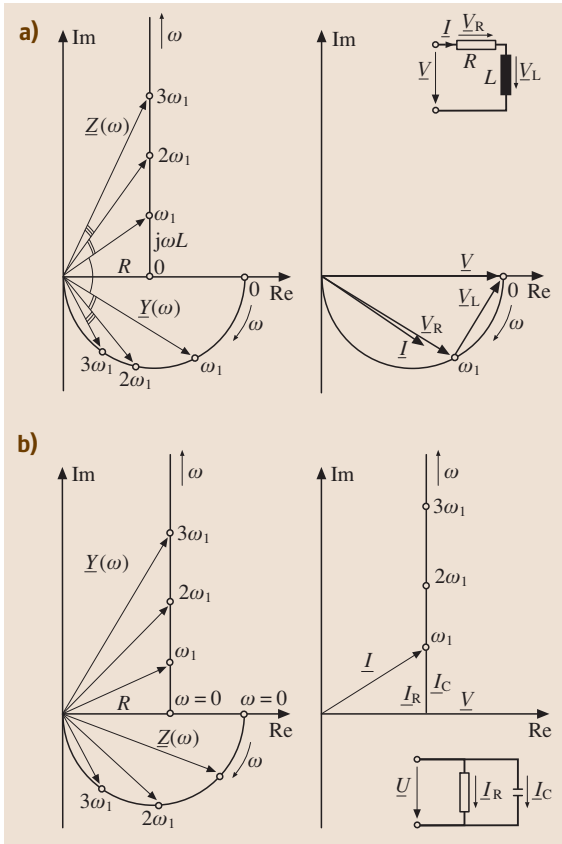


Fig. 17.11a,b Transfer locus diagram of an RL load: **(a)** serial circuit; **(b)** parallel circuit

The complex admittance of the circuit, which consists of the parallel connected resistance and inductance, can be seen in Fig. 17.11b.

$$\underline{Y} = \frac{1}{R} + \mathrm{j}\omega C \qquad (17.87)$$

More complicated circuits naturally create more highly organized transfer locus diagrams.

Oscillating Circuits and Filters

Passive equivalent circuits, which consist of capacitors and inductors, are oscillatory structures. Excitation between different energy storages (capacitors and inductors) could create interchange in the form of oscillations. A circuit is resonant at a certain frequency if the reactive components of the impedance (admittance) are equal to zero. Parallel and serial resonant circuits are two of the simplest resonant circuits. In the parallel circuit, it is assumed that current is the excitation and the

Table 17.2 Resonant circuit parameters

Circuit	Serial $\underline{Z} = R + \mathrm{j}(\omega L - \frac{1}{\omega C})$	Parallel $\underline{Y} = G + \mathrm{j}(\omega C - \frac{1}{\omega L})$
Resonance	$\omega_0 = \frac{1}{\sqrt{LC}}$, $Z_0 = \sqrt{\frac{L}{C}} = \frac{1}{Y_0}$	$\omega_0 = \frac{1}{\sqrt{LC}}$, $Z_0 = \sqrt{\frac{L}{C}} = \frac{1}{Y_0}$
Damping	$d_\mathrm{s} = \frac{1}{2}\frac{R}{Z_0} = \frac{1}{2Q_r}$ $\underline{Y} = Y_0 \frac{1}{2d_\mathrm{r} + \mathrm{j}\left(\frac{\omega}{\omega_0} - \frac{\omega_0}{\omega}\right)}$ $\underline{I} = \underline{Y}\,\underline{V}$	$d_\mathrm{p} = \frac{1}{2}\frac{G}{Y_0} = \frac{1}{2Q_\mathrm{p}}$ $\underline{Z} = Z_0 \frac{1}{2d_\mathrm{p} + \mathrm{j}\left(\frac{\omega}{\omega_0} - \frac{\omega_0}{\omega}\right)}$ $\underline{V} = \underline{Z}\,\underline{I}$

voltage the response, the converse being true in the serial case. Hence, the impedance of one can be obtained from the admittance of the other by interchanging L and C and replacing R in the serial circuit by G in the parallel circuit. Consequently, we can treat one of the circuits in detail and apply the results to the other one simply by interchanging the appropriate parameters.

Table 17.2 summarizes the most important parameters for resonant circuits, where d is the damping and Q is the performance.

For resonant circuits we typically consider the functions $\underline{Y}/\underline{Y}_0$ for serial and $\underline{Z}/\underline{Z}_0$ for parallel circuits.

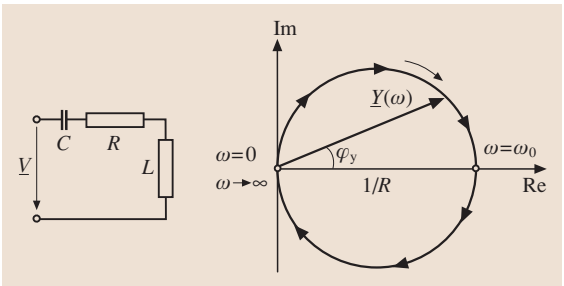


Fig. 17.12 Transfer locus diagram of a serial oscillating circuit

Four Terminals

An active four-terminal network consists of two terminal network input terminals (I_1, V_1) and two output terminals (I_2, V_2). An active four-terminal network consists of current or voltage sources, otherwise it is passive.

For passive terminal networks the following relations can by written:

Chain form (forward)

$$\begin{pmatrix} \underline{V}_1 \\ \underline{I}_1 \end{pmatrix} = \begin{pmatrix} \underline{A}_{11} & \underline{A}_{12} \\ \underline{A}_{21} & \underline{A}_{22} \end{pmatrix} \begin{pmatrix} \underline{V}_2 \\ \underline{I}_2 \end{pmatrix} \qquad (17.88)$$

Resistance form

$$\begin{pmatrix} \underline{V}_1 \\ \underline{V}_2 \end{pmatrix} = \begin{pmatrix} \underline{Z}_{11} & \underline{Z}_{12} \\ \underline{Z}_{21} & \underline{Z}_{22} \end{pmatrix} \begin{pmatrix} \underline{I}_1 \\ \underline{I}_2 \end{pmatrix}$$

Conductance form

$$\begin{pmatrix} \underline{I}_1 \\ \underline{I}_2 \end{pmatrix} = \begin{pmatrix} \underline{Y}_{11} & \underline{Y}_{12} \\ \underline{Y}_{21} & \underline{Y}_{22} \end{pmatrix} \begin{pmatrix} \underline{V}_1 \\ \underline{V}_2 \end{pmatrix}$$

Hybrid form

$$\begin{pmatrix} \underline{V}_1 \\ \underline{I}_2 \end{pmatrix} = \begin{pmatrix} \underline{H}_{11} & \underline{H}_{12} \\ \underline{H}_{21} & \underline{H}_{22} \end{pmatrix} \begin{pmatrix} \underline{I}_2 \\ \underline{V}_2 \end{pmatrix}$$

Normally, these coefficients can be determined by a short- and open-circuit investigation. The four-terminal network will be symmetrical if the resistance takes the form $Z_{22} = Z_{11}$ and $Z_{21} = Z_{12}$. The relations (17.88) are very useful for calculating more advanced circuits, e.g., filters. A passive terminal network can be characterized by two equivalent circuits: T or Π (Fig. 17.13).

For the T circuit using (17.89)

$$A_{11} = A = 1 - 2\omega^2 LC \qquad (17.89)$$

and $\omega_g = 1/\sqrt{LC}$.

The real impedance at the pass band will be defined as the characteristic impedance Z_W

$$Z_\mathrm{W} = \sqrt{\frac{A_{12}}{A_{21}}} = \sqrt{\frac{L}{C} - \omega^2 L^2}\,. \qquad (17.90)$$

Filters. Filters are circuits which are frequency selective [17.5]. By using various combinations of reactance and resistance, filter circuits can be made to either emphasize or deemphasize selected ranges of frequencies. There are four kinds of filters:

- Low-pass filters
- High-pass filters

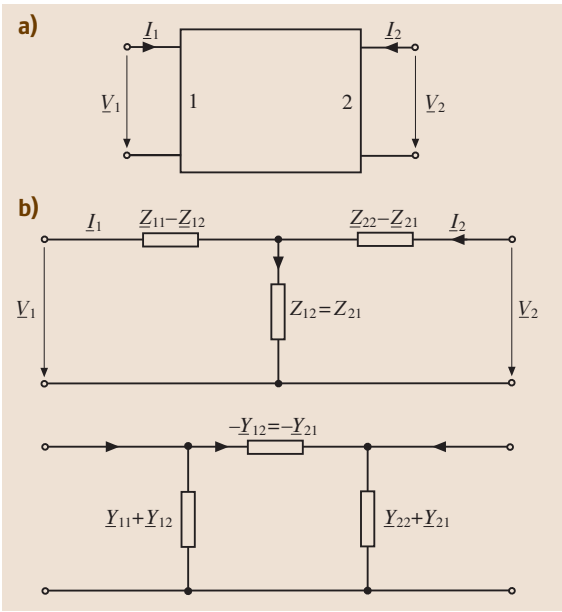


Fig. 17.13a,b Active four-terminal networks: **(a)** general circuit; **(b)** T and Π circuit [17.2]

- Resonant filters
- Multiple resonant filters

Properties of high-pass filters can be explained as follows. At first series RL and RC circuits are considered (Fig. 17.14). For the RC circuit in Fig. 17.14a the output is taken across the resistor for the high-pass filter. In this circuit, most of the input voltage drops across C at low frequencies (all of it at 0 Hz) because X_C is very large compared to R. As the frequency increases, X_C decreases so that less voltage is dropped across C and more across R. At very high frequencies X_C is so small compared to R that essentially all of the input voltage is dropped across R.

For the RL high-pass circuit shown in Fig. 17.14b, the output must be taken across the inductor. In this way, the output is zero at 0 Hz and increases as the frequency increases because of the larger voltage drop across X_L.

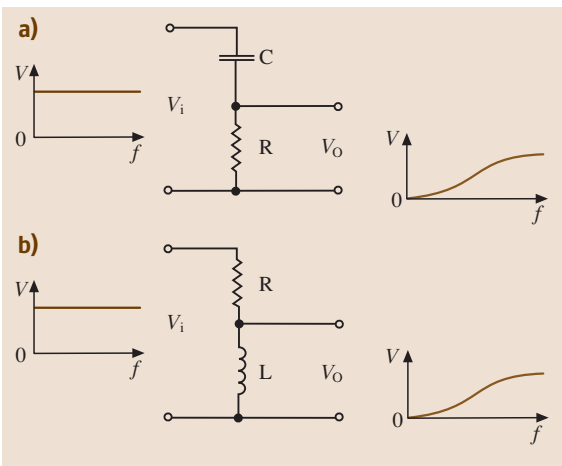


Fig. 17.14a,b High-pass filters: **(a)** series RC; **(b)** series RL

17.1.4 Networks

Transient Phenomena

Very often in networks transient phenomena take place, which consist of stationary state changes in passive elements. The states are characterized by currents or voltages at the clamps of the components and generally are excited by the switching processes taking place in the network.

The calculation of the transient phenomena can be carried out in the time domain. Moreover, for linear network systems, the Laplace transformation, which converts time functions into simple corresponding s-domain functions, can also be applied. For such systems, the operator equations resulting from a direct Laplace transformation can be created and solved by simple algebraic operations.

As mentioned, transient phenomena arise due to the closing and opening of the circuit switches. The state of the system before the switching occurs is described as the initial state. From general field theory it is known that the energy stored either in a magnetic field, in the case of an inductance, or in an electric field, in the case of a capacitor, cannot change its value the instant the switching takes place. From this law the transient phenomena in the network arise:

1. Calculation in time domain (state equation)
   The calculation of transient phenomena can be carried out by integration of the differential equations based on the set of Kirchhoff's laws for the given network. In a linear system with constant element parameters, the differential equations for the independent currents and voltages at the energy-storage components can be expressed as follows

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = \mathbf{A}x(t) + \mathbf{B}v(t)\,. \qquad (17.91)$$

   For a general description, $x(t)$ describes the state vector, $\mathbf{A}$ is interpreted as the state matrix, $\mathbf{B}$ is the

excitation matrix, and $v(t)$ represents the input signals which result from active elements in the circuit. The solutions to the above equations can be stated as shown in the following expression

$$x(t) = \mathrm{e}^{A(t-t_0)}x(t_0) + \int_{t_0}^{t} \mathrm{e}^{A(t-\tau)}\mathbf{B}v(\tau)\,\mathrm{d}\tau\ , \tag{17.92}$$

where $x(t_0)$ is the initial state vector, which disappears for zero initial conditions. The solution to this problem can also be expressed as the sum of a transient-state component and a steady-state component

$$x(t) = x_\mathrm{t}(t) + x_\mathrm{s}(t)\ . \tag{17.93}$$

The steady-state component $x_\mathrm{s}(t)$ can be solved by commonly used calculation procedures valid for steady-state AC and DC analysis. The transient element $x_\mathrm{t}(t)$ can be computed in this case with the help of a simplified formula

$$x_\mathrm{t}(t) = \mathrm{e}^{A(t-t_0)}x_\mathrm{t}(t_0)\ , \tag{17.94}$$

where the initial condition for the above equation is stated as the difference between the initial condition $x(t_0)$ and the initial condition for the steady-state component at the instant t_0

$$x_\mathrm{t}(t_0) = x(t_0) - x_\mathrm{s}(t_0)\ .$$

It can be seen that the solution of the transient phenomena is reduced to the calculation of the e^{At} component. In order to calculate this element, three main methods are used:

- The Sylvester equation
- The Cayley–Hamilton theorem
- Eigenvalue decomposition with Jordan form representation

Among these methods only the eigenvalue transformation will now be explained in an example. Assuming that the state matrix $\mathbf{A}$ is square, the eigenvalues for this matrix can be calculated with the help of the formula

$$\det(\mathbf{A} - \mathbf{I}s) = 0\ . \tag{17.95}$$

The calculated values s_k are the basis for obtaining the eigenvectors that build the transformation matrix $\mathbf{T}$. This matrix allows us to transfer the square matrix $\mathbf{A}$ into Jordan form $\mathbf{A}_\mathrm{j}$. The sought component e^{At} can be reformulated as

$$\mathrm{e}^{At} = \mathbf{T}^{-1}\mathrm{e}^{A_\mathrm{j}t}\mathbf{T}\ . \tag{17.96}$$

The elements of the Jordan matrix contain the linear combination of the exponential and polynomial function dependent on the multiplication factors and values of the eigenvalue of the state matrix $\mathbf{A}$.

2. Laplace Transformation

Transient phenomena in a network can be calculated with the help of the Laplace transformation as well. Using operator methods based on the Laplace transformation, the state equation can be written in the following form

$$(s\mathbf{I} - \mathbf{A})\boldsymbol{X}(s) - \boldsymbol{X}(t_0) = \boldsymbol{B}\boldsymbol{V}(s)\ , \tag{17.97}$$

where I is the identity matrix. The input signal is transformed into s-domain by the following integral

$$\mathrm{Im}[f(t)] = F(s) = \int_0^\infty f(t)\mathrm{e}^{-st}\,\mathrm{d}t\ . \tag{17.98}$$

After solving the much easier algebraic equation in s-domain, the solution needs to be transformed back from the s-domain into time domain

$$\mathrm{Im}^{-1}[F(s)] = f(t) = \int_{\sigma-j\infty}^{\sigma+j\infty} F(s)\mathrm{e}^{st}\,\mathrm{d}s\ . \tag{17.99}$$

The initial conditions for the state variables are considered by the Laplace transformation and are contained in the input signals (excitation signals). To obtain the Laplace function, correspondence tables can be a useful tool. The solution of the transient phenomena is obtained by the inversion of the square matrix $(\mathbf{I}s - \mathbf{A})$ and multiplication by its excitation vector $\boldsymbol{B}\boldsymbol{V}(s) + \boldsymbol{X}(t_0)$. As a result of this approach, the polynomial functions for the state variables are obtained and written as

$$x_i(s) = \frac{N_i(s)}{D_i(s)}\ . \tag{17.100}$$

Assuming that the denominator can be described by the polynomial function with stated poles

$$D_i(s) = (s - s_1)^{m_1}(s - s_2)^{m_2}\ldots(s - s_r)^{m_r} \tag{17.101}$$

and that the polynomial order equals

$$n = m_1 + m_2 + \ldots m_r\ , \tag{17.102}$$

the state i is represented as the following time function

$$x_i(t) = \sum_{k=1}^{r}\sum_{l=1}^{m_k} \frac{a_{kl}}{(m_k - l)!}t^{m_k - 1}\mathrm{e}^{s_k t}\ , \tag{17.103}$$

where the coefficients are evaluated from the expression

$$a_{kl} = \frac{1}{(l-1)!}\frac{\mathrm{d}^{l-1}}{\mathrm{d}s^{l-1}}\left[\frac{(s-s_k)^{m_k}N_i(s)}{D_i(s)}\right]\Bigg|_{s=s_k} . \tag{17.104}$$

The above approach is appropriate if the grade of the numerator is less than that of the denominator. Otherwise the numerator must be divided into two parts: polynomial parts for which the reverse Laplace transformation is the sum of the Dirac's function with the order equal to the grade individual polynomial component, and rational parts for which the above presented approach can be applied.

Selecting the analytical method suitable for finding the state variables of the given circuit depends on the number of variables, the input characters, and the network structure, but nowadays numerical methods based on the state equation are preferred.

3. Circuit with inductive and resistive load

If the network consists of a serial connection of a resistance, inductance, and voltage source (Fig. 17.15), the equation for such a structure can be written as

$$V_0(t) = V_\mathrm{R} + V_\mathrm{L} , \tag{17.105}$$

$$L_0\frac{\mathrm{d}i}{\mathrm{d}t} + R_0 i = v_0(t) . \tag{17.106}$$

For an assumed zero initial condition at the coil and a constant voltage supply $v_0(t) = V_0$, the above expression can be calculated with the help of the derivative common approach or Laplace transformation and written in the following form

$$i = \frac{V_0}{R_0}\left(1 - \mathrm{e}^{-t/T}\right), \quad T = \frac{L_0}{R_0} . \tag{17.107}$$

The response of the current increases due to the exponential function (Fig. 17.16a) with a time constant $T = L_0/R_0$. When excitation is an AC (alternating) voltage of the form

$$v(t) = \sqrt{2}V\cos(\omega t + \varphi) , \tag{17.108}$$

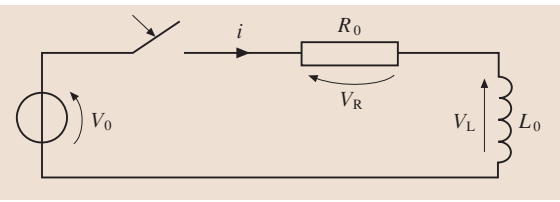


Fig. 17.15 Electric circuit of resistance R_0 and inductance L_0 in series

the solution of the derivative equation can be formulated

$$i(t) = \sqrt{2}I\left[\cos(\omega t + \varphi - \phi) - \mathrm{e}^{-t/T}\cos(\varphi - \phi)\right], \tag{17.109}$$

where Z is the impedance module, described as $\sqrt{R_0^2 + \omega^2 L_0^2}$, $I = V/Z$ is the effective (RMS) current value, and ϕ is the shift angle defined as $\phi = \arctan\omega T$ dependent on the angular speed of the AC voltage and the time constant. The transient state in this case is characterized by both the oscillatory state resulting from the sinus waveform excitation and the instantaneous state, which depends on the moment of switching the source (Fig. 17.16b). This undesirable instantaneous state has a maximum value for $\varphi = \phi$ and disappears when the condition $|\varphi - \phi| = \pi/2$ is reached. If the coil has a significant inductance compared to the resistance $R \ll \omega L$ and is switched on at the maximum of the excitation voltage, the current instantaneously reaches

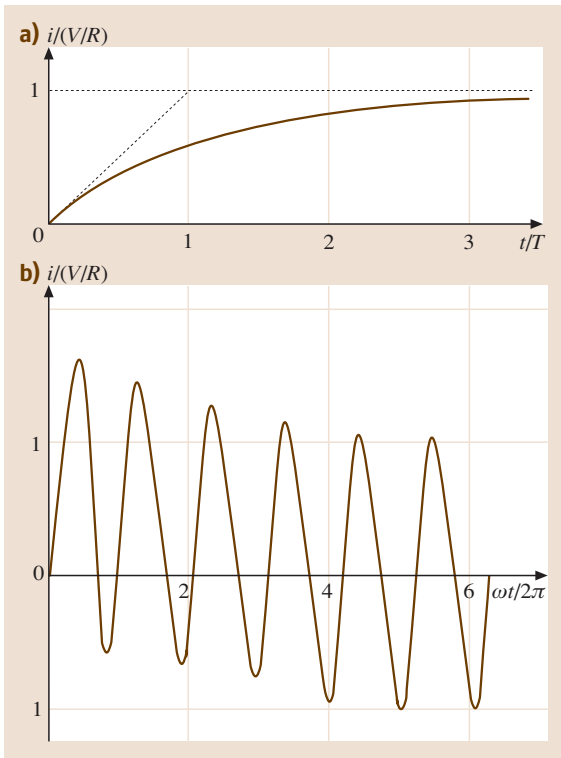


Fig. 17.16a,b Switch processes by resistance–inductance load: **(a)** at a DC voltage source; **(b)** at an AC voltage source

the steady state with the angular speed of the voltage, but if the source is switched on during the zero crossing of the voltage, the transient current can achieve twice the magnitude of the stationary state.

4. Circuit with a resonance branch
When we consider the serial connection of a constant voltage source, a resistor, a coil, and a capacitor with zero initial conditions presented in Fig. 17.17, the equation for such a framework can be defined as

$$L_0 \frac{\mathrm{d}i}{\mathrm{d}t} + R_0 i + \frac{1}{C_0} \int i \, \mathrm{d}t = V_0 \,. \tag{17.110}$$

The solution of this equation in the time domain is represented by the linear combination of two exponential functions in the form

$$i = A_1 \mathrm{e}^{p_1 t} + A_2 \mathrm{e}^{p_2 t} \,. \tag{17.111}$$

To find the parameters p_1 and p_2, the second-order characteristic equation resulting from the derivative approach has to be calculated

$$p^2 + 2\delta p + \omega_0^2 = 0 \,, \tag{17.112}$$

where $\omega_0^2 = 1/(L_0 C_0)$ is the resonant frequency and $\delta = R_0/2L_0$ is the damping factor.
The expression above has two roots

$$p_{1,2} = -\delta \pm \sqrt{\delta^2 - \omega_0^2} \,. \tag{17.113}$$

Due to the different characteristic values, which depend on the damping factor and the resonant frequency, we can distinguish three main cases:

– Overdamped circuit. When the element values are such that $\delta^2 > \omega_0^2$ here are two real roots that give the following response

$$i = \frac{V_0}{\alpha L_0} \mathrm{e}^{-\delta t} \sinh(\alpha t) \,, \tag{17.114}$$

where $\alpha = \sqrt{\delta^2 - \omega_0^2}$.

– Critically damped circuit. This occurs when the element values exactly satisfy the condition $\delta^2 = \omega_0^2$. In this case we say that the roots are repeated. The expression to solve such a problem is written

$$i = \frac{V_0}{L_0} t \mathrm{e}^{-\delta t} \,. \tag{17.115}$$

– Underdamped circuit. When the characteristic equation yields complex roots, we say that the circuit is underdamped because its natural response exhibits an oscillatory behavior. The solution for this case is expressed as

$$i = \frac{V_0}{\omega L_0} \mathrm{e}^{-\delta t} \sin(\omega t) \,, \tag{17.116}$$

where $\omega = \sqrt{\omega_0^2 - \delta^2}$ is the damped frequency. The current characteristics for all three cases are presented in Fig. 17.18. The solution for nonzero initial conditions is more complicated and contains additional components resulting from the energy stored in the inductive and capacitive elements.
In addition, in the case of an AC instead of DC source, we obtain state variables that are a combination of the given harmonic excitation from the source and instantaneous responses.

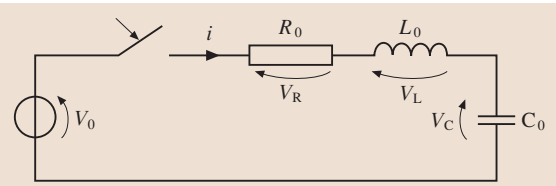


Fig. 17.17 Electric circuit with a resistance R_0, impedance L_0, and capacitance C_0 in series

Network Analysis

Circuit theory includes all issues dealing with the dynamic and stationary behavior of the network. The principles for the calculation of the behavior of the circuit are based on the two fundamental Kirchhoff's laws:

- Kirchhoff's current law (KCL), which says that, for all times t, the algebraic sum of the currents flowing toward a node and the currents flowing away from that node is equal to zero (17.25), and
- Kirchhoff's voltage law (KVL), which says that, for all closed node sequences (meshes) and for all

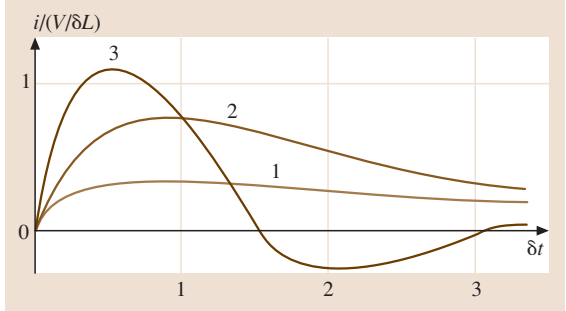


Fig. 17.18 Pure resonance circuit switching: 1 – overdamped circuit, 2 – critically damped circuit, and 3 – underdamped circuit

times t, the algebraic sum of all node-to-node voltages around the chosen closed node sequence is equal to zero (17.26)

Network analysis is based on the above laws. The first Kirchhoff law is used for node analysis and the second is applied to mesh analysis.

To study wide circuit systems, a topological approach can be used to facilitate the network calculation. This allows one to build the equation systems for the given circuit and to compute them with the help of numeric techniques.

The interconnection properties of the circuit can be best defined by a circuit graph. A circuit graph is specified by a set of nodes together with a set of branches. If for each branch an orientation is given, indicated by an arrow on the branch, we call the graph a directed graph or digraph. Only directed graphs have practical meaning for circuit calculation. The fundamental terms for network analysis are nodes, branches, meshes, and trees. As an illustrative example the circuit in Fig. 17.3 can be assigned a graph with $b = 6$ branches and $k = 4$ nodes. Figure 17.3 has been slightly modified as shown in Fig. 17.19.

Based on this graph we can conclude that the circuit results in $p = k - 1$ independent equations and that the fundamental node matrix, commonly called the admittance matrix, created for this circuit has the following form

$$G = \begin{pmatrix} G_1 + G_2 & -G_1 & 0 \\ -G_1 & G_1 + G_3 + G_5 & -G_3 \\ 0 & -G_3 & G_3 + G_4 \end{pmatrix} . \tag{17.117}$$

This matrix is symmetrical, which is always the case when the circuit does not have controlled active elements. The admittance matrix is created in the following way:

- Main diagonal components $G_{i,i}$: the positive sums of all the conductances and *susceptances* which are connected to the node i, and
- Nondiagonal components $G_{i,j}$: diagonal components that reflect the mutual negative conductance between connected nodes i and j

The number of independent equations m in mesh analysis can be calculated from the formula $m = b - k + 1$. In order to create the equations for mesh current studies, the tree of the circuit must first be assigned. This term plays a central role in graph theory. A tree is a subgraph of the connected digraph and satisfies the following three fundamental properties:

- It is connected.
- It contains all nodes of the circuit.
- It has no loops.

The branches, which create the loops with the tree branches, assign the mesh current and are called chords or co-trees. In each equation only one chord and arbitrary tree paths are taken into account. The above mentioned rules can be used in writing down the mesh impedance matrix. For Fig. 17.20 this matrix is expressed as

$$R = \begin{pmatrix} R_1 + R_3 & -R_1 & -R_3 \\ -R_1 & R_1 + R_2 + R_3 & -R_5 \\ -R_3 & -R_5 & R_3 + R_4 + R_5 \end{pmatrix} . \tag{17.118}$$

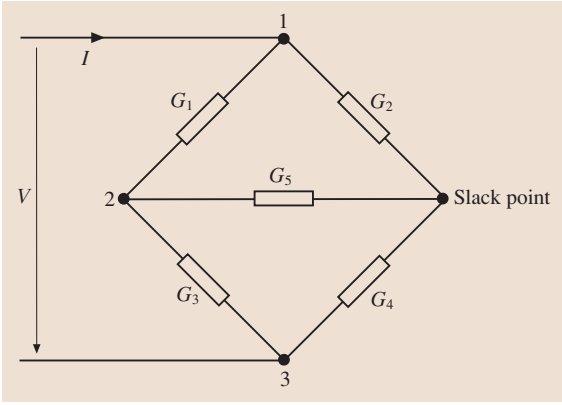


Fig. 17.19 Bridge circuit with four nodes and six branches

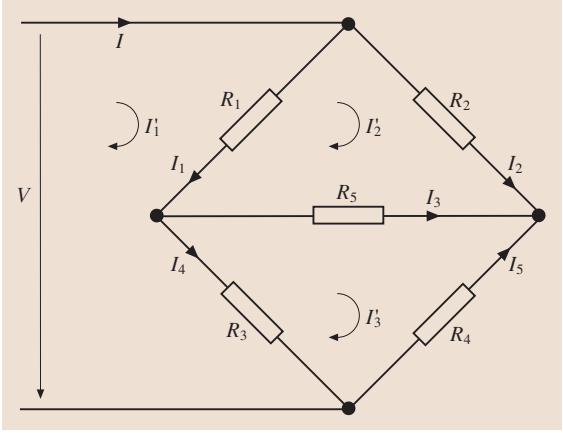


Fig. 17.20 Bridge circuit with I'_1, I'_2, and I'_3 mesh currents

It is dependent on the assumed mesh current directions (see I_1', I_2', and I_3', in Fig. 17.20).

The diagonal components of the mesh impedance matrix contain resistances that appear in the loops created by tree branches and chords. The nondiagonal elements of the matrix are built from the mutual resistances of the assumed tree branches.

The presented matrices provide the basis to calculate the currents and voltages at the clamps of the load. For practical reasons, node analysis is more often used in calculations due to the simple construction procedure. For mesh analysis, the allocation of arrows to the branches plays a very important role, and becomes complicated for large systems.

17.1.5 Materials and Components

Conductors, Semiconductors, and Insulators

The specific resistance for the various materials used in electrical components can range from 10^{-6} to 10^{20} Ω cm (Fig. 17.21) [17.6]. The basis for current conduction is electrons and p-hole mobility. Based on their different carrier types and carrier mobility, three main groups of materials can be distinguished:

- Metallic conductors (especially Cu, Al, Ag)
- Semiconductors (Si, Ge, Se, as well as compounds of the third and fifth groups of the periodic table)
- Insulators (organic and inorganic compounds such as porcelain, glass, and mica)

The semiconductors are elementary elements for electronic and power electronic circuits [17.7]. If a dice (junction) has an undisturbed and pure structure than at low temperatures it does not conduct any carriers and behaves as an insulator. Carrier mobility can be induced by heat or light irradiation.

Because of different dopants in the semiconductors a distinct group of carrier types appear. When the acceptors (the third column of the periodic table) dope the donors (the fifth column of the periodic table), various p- and n-carriers arise.

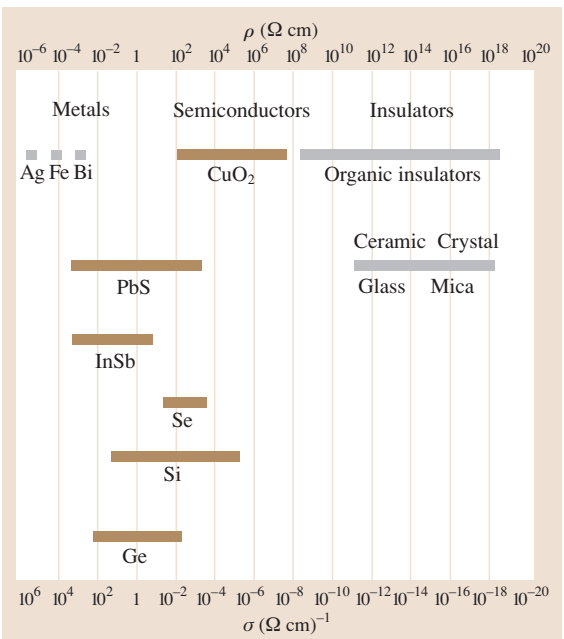


Fig. 17.21 Specific resistance for electrical materials

Special Properties of Conductors

Superconductor. The behavior of metallic conductors is dependent on temperature. In an ordinary working range of the conductor wires exhibit linear dependency. In the low-temperature range some metals and metal alloys possess superconductive properties. When approaching the so-called absolute zero temperature, at the temperature T_c the resistance vanishes completely.

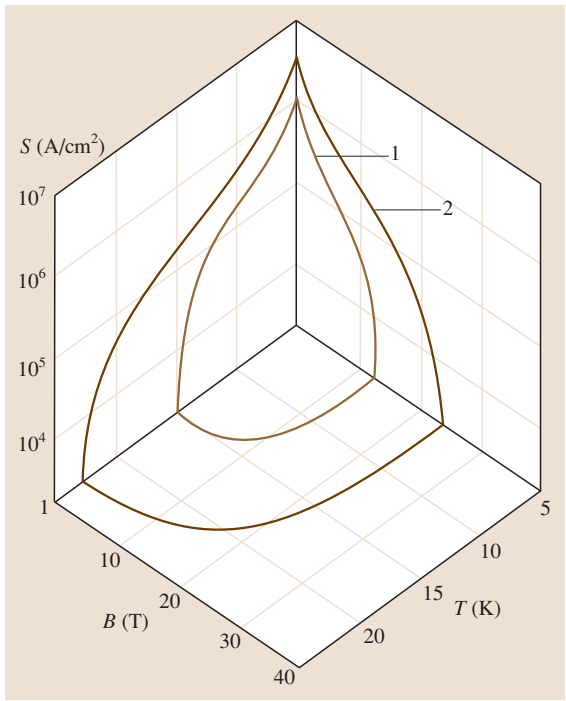


Fig. 17.22 Jump temperature of superconductors: 1 – NbTi, 2 – Nb_3Sn

In order to keep the superconductive effect, certain critical values of the current density and external magnetic field cannot be exceeded (Fig. 17.22). Superconductors can be used in energy engineering for electrical machines, distribution feeders, and storage systems. For this purpose, high-temperature superconductors have been designed based on compounds of NbTi ($T_c = 9.3\,\mathrm{K}$), Nb_3Sn ($T_c = 18.0\,\mathrm{K}$) and V_3Ga. For example, superconductance for a Ni and Ti compound is achieved at a temperature of 4.2 K, a current density of $70\,\mathrm{kA/cm^2}$, and a critical magnetic flux density of 8 T. Due to the low temperature that must be maintained for superconduction, adequate cooling materials – most often liquid helium (LHe) and liquid nitrogen (LN_2) – and appliances are required. Recently, new technologies have been being developed to obtain high-temperature superconductors in which the superconductive properties appear at temperatures of about 100 K. The compound $YBa_2Cu_3O_7$, for example, exhibits the zero-resistance effect at 93 K. High-temperature superconductors have the advantage of lower cost and higher efficiency. For example, in order to achieve the superconducting temperature for such materials, cheaper nitrogen instead of high-priced helium can be used. Additionally less energy is needed to maintain the appropriate conditions.

However, due to their low critical current density and some problems with the production of superconductor components, their practical application still lies in future.

Hall Effect. When current flows through a metal strip with a rectangular cross-section a magnetic field perpendicular to it generates a voltage between the two opposite margins of that metal strip; this is known as the Hall voltage. This Hall effect can serve to measure the magnetic field. Furthermore, based on this effect the currents in wires can be measured with the help of special equipment.

In order to use the Hall effect for measurements, a thin flat blank with a high Hall factor is applied. For example, a Hall-effect probe made of InAs materials achieves a Hall voltage of about 100 mV at currents of about 0.1 A and a magnetic induction of 1 T.

Seebeck and Peltier Effects. Two further effects are the Seebeck and Peltier effects which both concern the thermovoltage. A thermovoltage is produced by two wires connected to each other as a thermocouple. In a closed circuit loop the thermovoltage is noticeable when the two brazed points have different temperatures. The Seebeck effect states that the generated voltage between two different metal wires connected in one circuit is proportional to the difference of temperatures between the two thermo-elements. One of these joints is usually kept at a fixed reference temperature, for instance in a container filled with water and ice (for a reference temperature of 0 °C). When a copper wire is connected to a constantan wire in one circuit and a temperature difference of 100 K exists, a thermovoltage of 4.15 mV is produced. A second effect, known as the Peltier effect, is the inverse of the Seebeck effect. Current flowing through a thermocouple irrespectively to Joule heat feeds the one brazed point with heat which is dissipated from the other brazed point. The Peltier thermal effect is proportional to the current flowing in the circuit. This phenomenon can be found in various circuit elements subjected to cooling.

Materials in an Electric Field

Insulation materials are characterized by their relative permittivity ε_r and electric breakdown strength E_d. In an electric field the molecules are polarized by the electric charge. The electric field strength is combined with the polarization by a formula

$$P = \chi_e \varepsilon_0 E\,, \tag{17.119}$$

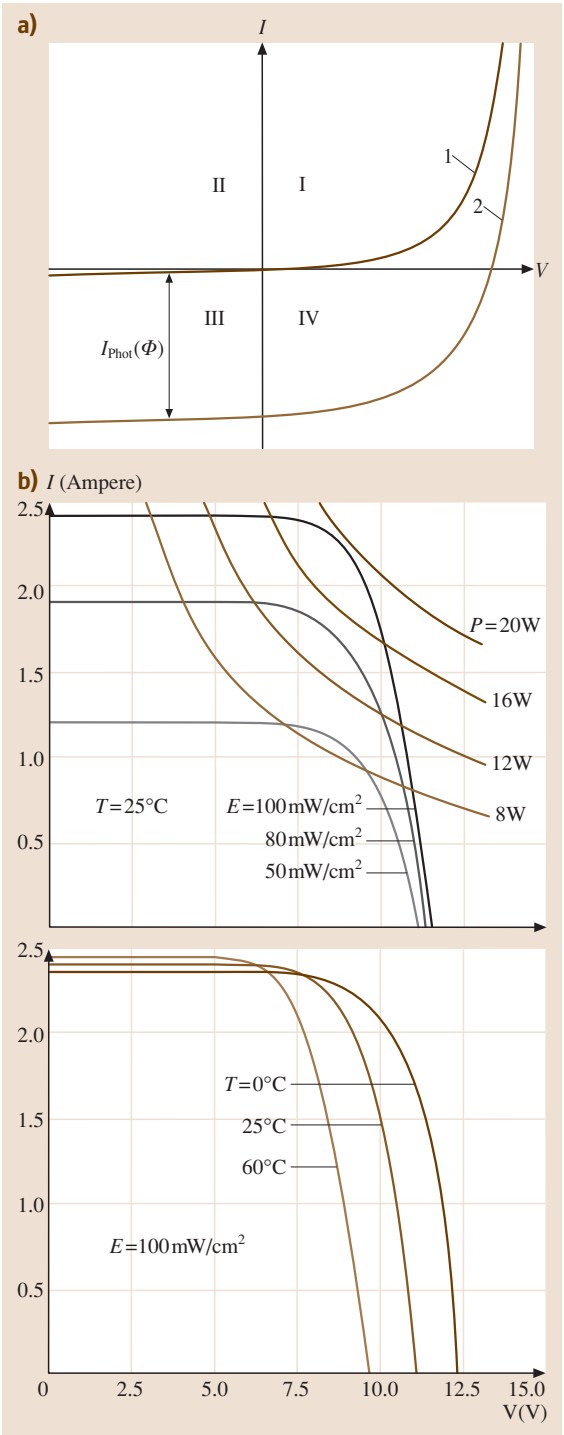
where $\chi_e = \varepsilon_1 - 1$ is the dielectric susceptibility.

Some dielectric materials have a nonlinear relationship between polarization and electric field strength and show ambiguous hysteresis properties. This behavior is called ferroelectricity.

Piezoelectricity. Some crystals can become polarized by applying pressure or tensile stress. On the opposite edges of the crystal surface charges arise with different signs. Reciprocally, for these materials it is possible to apply an electrical field which causes a change in the length of the material. This change is dependent on the polarization and direction of that field.

In general, piezoelectric materials are used for the conversion of electromechanical oscillations. For example, piezoelectric transducers are applied in measurement techniques, microphones, and especially in quartz clocks. The most recent application for piezoelectric transducers is in drive devices and in the field of noise reduction.

Photovoltaic Solar Cells. Solar cells are photovoltaic (PV) elements with a p–n junction in which light irradiation causes the separation between the electrons and p-holes and a voltage can develop [17.8]. The energy



generated can be used to supply a load connected to the photovoltaic element. The behavior of the photovoltaic elements is described by its diode characteristic, which is dependent on light irradiation; therefore the negative photocurrent is shifted with respect to the real diode curve (Fig. 17.23a). A photovoltaic element works in the fourth quadrant between the short-circuit current and the open-circuit voltage.

Solar cells can be produced from mono- or polycrystalline and amorphous silicon. Silicon cells based on different technologies differ in terms of their efficiency factor and expenditure in manufacturing; commercial cells have an efficiency value of 10–20%. Through serial and parallel connection of these cells, solar modules are obtained. Figure 17.23b illustrates the characteristics of a solar module that contains 5×4 multicrystalline cells with a $10\times10\,\mathrm{cm}^2$ surface area. The standard test conditions (STC) for solar modules are a module temperature of 25 °C, light irradiation of $100\,\mathrm{mW/cm}^2$, and a light spectrum of AM = 1.5. This system generates about 19.2 W of power at its maximum-power point (MPP).

Solar generators are used as the main energy source in satellite technology. Due to its renewable and environmentally friendly properties these generators are beginning to play a very important role in energy technology.

Materials in a Magnetic Field

The magnetic properties of materials are defined by the magnetic susceptibility χ_m, which describes the connection between the magnetization M and magnetic field intensity H

$$M = \chi_\mathrm{m} H\,; \quad \chi_\mathrm{m} = \mu_\mathrm{r} - 1. \tag{17.120}$$

The following material groups can be distinguished:

- Paramagnetic materials, where the μ_r called relative magnetic permeability value is slightly greater than 1 (e.g., $\chi_\mathrm{m} = 0.21\times10^{-4}$ for Al)
- Diamagnetic materials, where the μ_r value is slightly smaller than 1 ($\chi_\mathrm{m} = -0.19\times10^{-4}$ for Ag)
- Ferromagnetic materials, where the χ_m value is much greater than 1 and can be up to 1×10^5 (e.g., Fe, Ni, Co, and some alloys)

Fig. 17.23a,b Characteristics of a PV converter. **(a)** Photo element as diode with photocurrent, 1 – traditional diode, 2 – solar cell; **(b)** solar module (AEG PQ 10/20)◀

The magnetization characteristics $B(H)$ or $M(H)$ of ferromagnetic materials exhibit saturation and hysteresis properties. If this material is excited by a pulsating field then magnetization losses arise, composed of eddy currents and hysteresis losses.

Ferromagnetic materials in an external magnetic field exhibit elastic length changes. This phenomenon is called magnetostriction and is utilized in the generation of ultrasonic oscillations. The magnetostriction phenomenon is also responsible for unpleasant noise from transformers.

Electrolytic Charge Transfer

The charge transport of currents through electrolytes (bases, acids, salines and fused salts) follows the ions, namely the positive cations and negative anions. Negative ions in a fluid conductor migrate from cathode to anode while positive ions migrate in the opposite direction under the influence of an electric field. This phenomenon is accompanied by material transport, and its effect is utilized technically in electrolysis.

Electrodeposition is the use of electrolysis to separate materials. This technique makes it possible to achieve a high level of cleanliness of the separated materials. For example copper obtained from the electrolysis process exhibits a cleanness of about 99.9%. Aluminum produced by electrodeposition involves the additional materials of bauxite and cryolite in fluid state. Electroplating is the production of metallic coatings using an electrolysis procedure (for example, nickel plating).

17.2 Transformers

17.2.1 Single-Phase Transformers

Working Principle and Equivalent Circuit Diagram

A simple transformer device has two magnetically coupled windings [17.9]. One side of the transformer is called the primary winding and the other side is the secondary winding. The transformer consists of copper windings and an iron core carrying flux which produces magnetic coupling between the two windings of the transformer (Fig. 17.24). Depending on construction principles, usage, power rating, and other factors we can distinguish between the following groups and subgroups:

- Usage: power, regulating, generator, and transmission transformers
- Power rating: small, medium, and large power transformers, and high-performance transformers
- Type of cooling: oil-immersed and dry-type transformers
- Core type: core- and shell-type transformers
- Transformation ratio: step-up and step-down transformers
- Arrangement of windings: separate winding transformers and auto-transformers
- Number of phases: single- and three-phase transformers, grouped into three- and five-limb cores

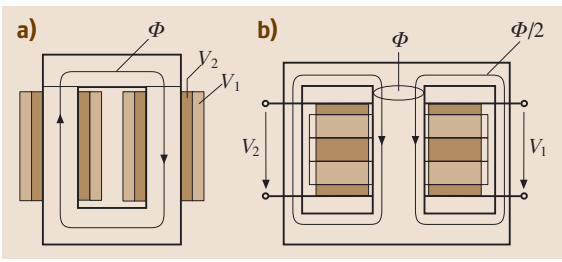


Fig. 17.24a,b Single-phase transformer construction: **(a)** UI transformer; **(b)** EI transformer

The transformer can be mathematically analyzed as a reverse two-port network. The magnetic properties are characterized by the inductivity of the windings, L_1 and L_2, and the mutual inductivity M. Currents flowing through the windings induct primary and secondary flux that are combined with each other as written below

$$\Psi_1 = L_1 i_1 + M i_2 \,,$$
$$\Psi_2 = L_2 i_2 + M i_1 \,. \qquad (17.121)$$

For the transformer a coefficient k is defined to describe the coupling

$$k = \frac{M}{\sqrt{L_1 L_2}} \,, \quad 0 \le k \le 1 \,. \qquad (17.122)$$

The coupling coefficient k of a physical transformer is determined by a number of factors: the magnetic properties of the core on which the primary and secondary coils are wound, the number of turns of each coil, the coils' relative positions and their physical dimensions. If k is close to zero, we say that the coils are

loosely coupled, as is the case for air-core transformers. If k is close to one, we say they are tightly coupled as occurs for iron-core transformers.

The windings in a real transformer also have resistances, so that, for the assumptions above, the following pair of equations in the time domain can be formulated

$$v_1 = R_1 i_1 + L_1 \frac{\mathrm{d}i_1}{\mathrm{d}t} + M \frac{\mathrm{d}i_2}{\mathrm{d}t} ,$$
$$v_2 = R_2 i_2 + L_2 \frac{\mathrm{d}i_2}{\mathrm{d}t} + M \frac{\mathrm{d}i_1}{\mathrm{d}t} . \qquad (17.123)$$

In Fig. 17.25a a schematic transformer that consists of resistances, self-inductances, and mutual inductance is presented for which the above expressions apply. This formulation can be used for steady-state AC analysis of the system, after being simplified to

$$\underline{V}_1 = (R_1 + \mathrm{j}\omega L_1)\underline{I}_1 + \mathrm{j}\omega M \underline{I}_2 ,$$
$$\underline{V}_2 = (R_2 + \mathrm{j}\omega L_2)\underline{I}_2 + \mathrm{j}\omega M \underline{I}_1 , \qquad (17.124)$$

where ω is the angular speed of the AC excitation.

In order to simplify the AC circuit calculations when transformer devices are used, the turn ratio is introduced

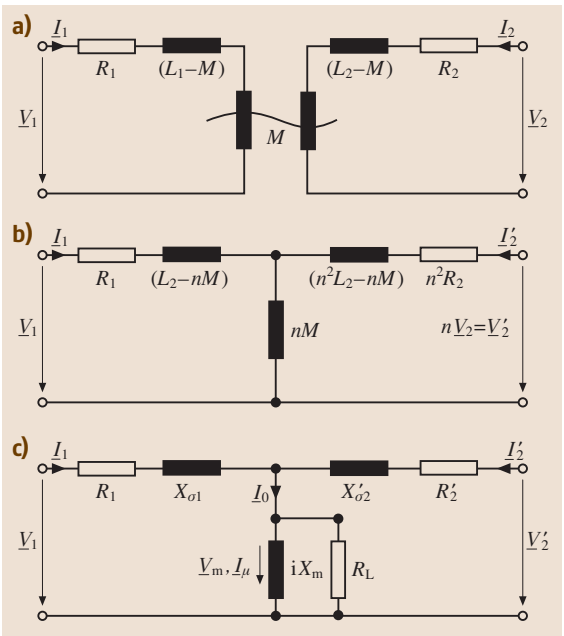


Fig. 17.25a–c Transformer schema with two windings: **(a)** principle circuit; **(b)** conversion of secondary side into primary side; **(c)** circuit for AC analyses [17.2]

and denoted as n

$$n = \frac{N_1}{N_2} . \qquad (17.125)$$

Using this ratio the transformer can be interpreted as one galvanically coupled circuit, forming a T-equivalent network (Fig. 17.25b,c). Therefore, the quantities of the secondary side have to be recalculated in order to be based on the primary side using the following equations

$$\underline{V}'_2 = n\underline{V}_2 , \quad \underline{I}'_2 = \underline{I}_2/n \qquad (17.126)$$

and, for passive elements the following connections are fulfilled

$$\underline{R}'_2 = n^2 R_2 , \quad L'_2 = n^2 L_2 . \qquad (17.127)$$

These relations obey the invariant-power condition for this two-port network study. The introduced turn ratio coefficient allows us to identify the simple connection between the main flux Φ_m that reflects current generated in the cross branch and the leakage fluxes Φ_1 and Φ_2 induced by the primary and secondary windings of the transformer.

The inductive transformer elements appear as the main inductivity L_m and the leakage inductivities $L_{\sigma 1}$ and $L_{\sigma 2}$. These parameters are reflected by the following relation

$$L_\mathrm{m} = nM , \quad L_{\sigma 1} = L_1 - L_\mathrm{m} ,$$
$$L_{\sigma 2} = L_2 - M/n . \qquad (17.128)$$

For transformer analysis it is very important to take into account not only the winding losses but also the magnetizing losses in the transformer core. This can be done by an additional parallel resistance R_L connection to the main inductivity in the T-network schema. The losses in the core result from eddy currents and magnetization of the hysteresis loop. The hysteresis losses are proportional to the working frequency and the eddy-current losses are proportional to the square of the working frequency. If a transformer works with fundamental frequency $f = \omega/2\pi$ then all of the circuit parameters can be described in complex numbers with assigned reactances as shown in Fig. 17.25c.

Voltage Induction

A current i flowing through a winding of the transformer core produces a magnetic field around the core.

If the excited current is time varying then it generates a time-varying voltage that, when transferred to the primary side of the transformer, takes the following form

$$V_\mathrm{m} = w_1 \frac{\mathrm{d}\Phi_\mathrm{m}}{\mathrm{d}t} = -\frac{\mathrm{d}\Psi_\mathrm{m}}{\mathrm{d}t} \,. \tag{17.129}$$

If the flux changes to a sine function then the inducted voltage will also have a sine waveform with the same frequency as the drawn current. The effective (RMS) value of the voltage is described as

$$V_\mathrm{m} = \frac{\omega}{\sqrt{2}} w_1 \Phi_\mathrm{m} = 4.44 f w_1 \hat{B} A_\mathrm{Fe} \,. \tag{17.130}$$

The generated voltage is proportional to the working frequency f, the winding number w_1, the magnetic induction magnitude $\hat{B}$, and the cross-section A_Fe.

No-Load and Short-Circuit Conditions

When the secondary winding of the transformer works in the open circuit, this means that $\underline{I}_2 = 0$, and the voltage at the output clamps of transformer is approximated by the formula

$$\underline{V}_{20} \approx \underline{V}_1 / n \,. \tag{17.131}$$

In this case, the current $\underline{I}_0$ in the primary winding is moved back by about 90° with respect to the supply voltage. The active power for such a working condition of the transformer reflects mainly the eddy current and hysteresis losses due to magnetization of the core, and the reactive power results from the almost purely inductive character of the device. The relations for these working conditions are expressed in the following form

$$P = V_1 I_0 \cos\varphi_0 \,, \quad Q = V_1 I_0 \sin\varphi_0 \tag{17.132}$$

for the active and reactive power, respectively.

The characteristic value for the transformer is an open-circuit current, which is defined as the ratio of the current I_0 in the open-circuit condition to the nominal current of the transformer

$$i_0 = \frac{I_0}{I_\mathrm{N}} \,. \tag{17.133}$$

This factor amounts to only a few percent. The P and Q quantities in the open-circuit condition serve as the basis for calculating the mentioned parameter R_L, which reflects the losses in the core, and X_m, which describes the main inductance in the transformer circuit.

The no-load characteristics $V_1 = f(I_0)$ exhibit the saturation features of the transformer, which must be taken into account during transformer design.

If a transformer works in short-circuit conditions of the secondary winding the following conditions are fulfilled

$$\underline{V}_2 = 0 \,, \quad \underline{I}_{2s} \approx -n\underline{I}_1 \,. \tag{17.134}$$

The flux produced by the current flowing in the primary winding is almost completely compensated by the flux produced by the current in the secondary winding.

The wire resistances and leakage reactance can be roughly obtained with the help of the following expressions

$$\begin{aligned} &\underline{Z}_\mathrm{s} = R_\mathrm{s} + \mathrm{j}X_\mathrm{s} \,, \quad R_\mathrm{s} = R_1 + n^2 R_2 \,, \\ &X_\mathrm{s} = \omega\left(L_{\sigma 1} + n^2 L_{\sigma 2}\right) \,. \end{aligned} \tag{17.135}$$

Figure 17.26 illustrates a further connection between the quantities in graphical form and introduces the term relative short-circuit voltage, which is defined as

$$v_\mathrm{s} = X_\mathrm{s} \underline{I}_\mathrm{N} / \underline{V}_\mathrm{N} \,. \tag{17.136}$$

For a power transformer these values can be up to 6%.

Phasor Diagram

In order to better understand the phenomena arising in the transformer, a phasor diagram is often used. In such a case the secondary-side quantities are transferred to the primary winding of the transformer and the relationship between currents and voltages at the clamps of the transformer are as follows

$$\begin{aligned} &\underline{V}_1 = (R_1 + \mathrm{j}\omega L_\mathrm{s})\underline{I}_1 + \mathrm{j}\omega L_\mathrm{m}\underline{I}_\mu \,, \\ &\underline{V}_2 = \mathrm{j}\omega L_\mathrm{m}\underline{I}_\mu + (R'_2 + \mathrm{j}\omega L'_{\sigma 2})\underline{I}'_2 \,, \end{aligned} \tag{17.137}$$

where $\underline{I}_\mu = \underline{I}_1 + \underline{I}'_2$ is defined as a magnetization current. If the additional core losses represented by the

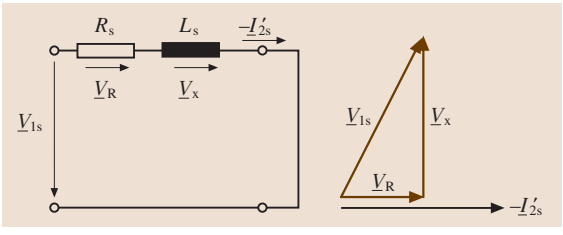


Fig. 17.26 Schema of a phasor diagram in short circuit

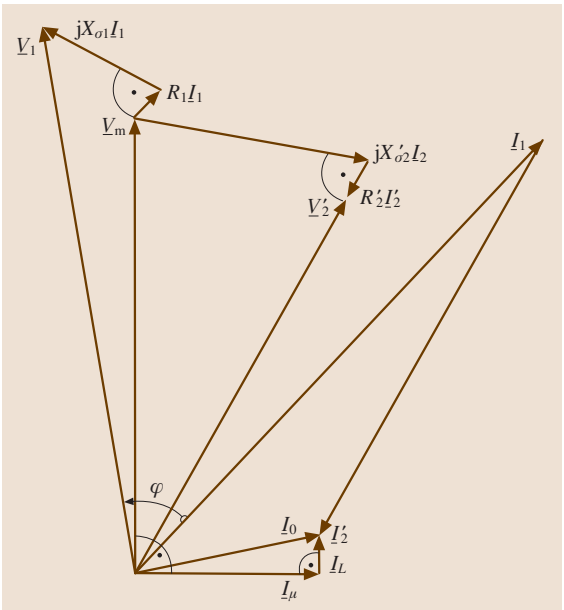


Fig. 17.27 Phasor diagram of transformer operation at resistance–inductance load [17.2]

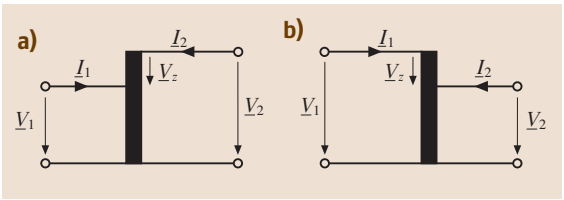


Fig. 17.28a,b Auto-transformer: **(a)** step-up; **(b)** step-down

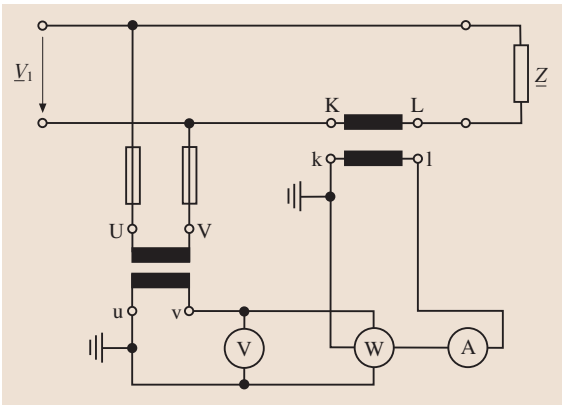


Fig. 17.29 Instrument transformer for voltage and current in a single-phase circuit

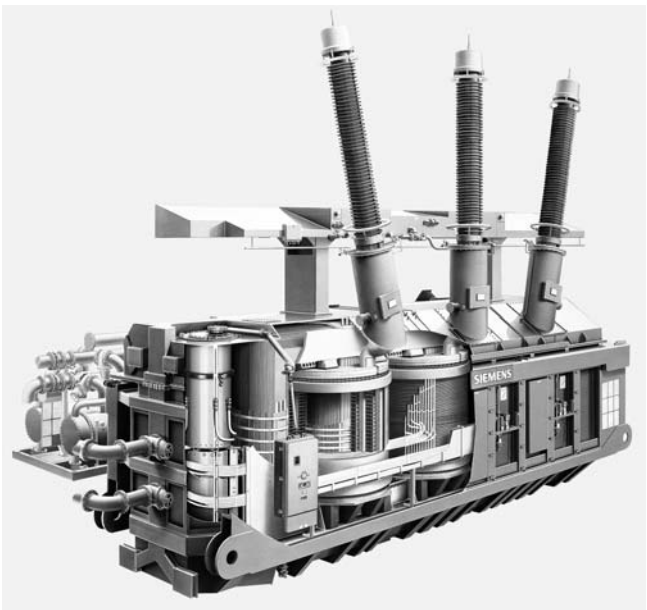


Fig. 17.30 Three-phase power transformer for public energy supply > 150 MVA (SIEMENS)

resistance R_L are taken into account, the behaviors of the currents and voltages at the transformer are described by the phasor diagram in Fig. 17.27.

It can be seen that the magnetization current is shifted back by 90° with respect to the main voltage, denoted by $\underline{V}_m = j\omega L_m \underline{I}_\mu$. The whole current $\underline{I}_0$ characterizes properties of the main flux in the transformer and can be divided into two geometric components: the primary and the related secondary current.

Transformer are called auto-transformer if the primary windings is part of the secondary windings or vice versa. The windings of the auto-transformer do not have galvanic separation, as shown in Fig. 17.28. Auto-transformers have lower leakage reactances, lower losses, and smaller exciting current when the voltage ratio does not differ too greatly from 1 : 1. The disad-

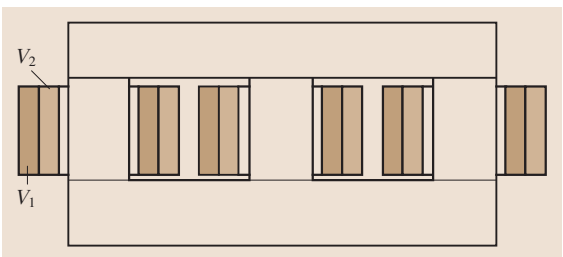


Fig. 17.31 Three-phase transformer construction

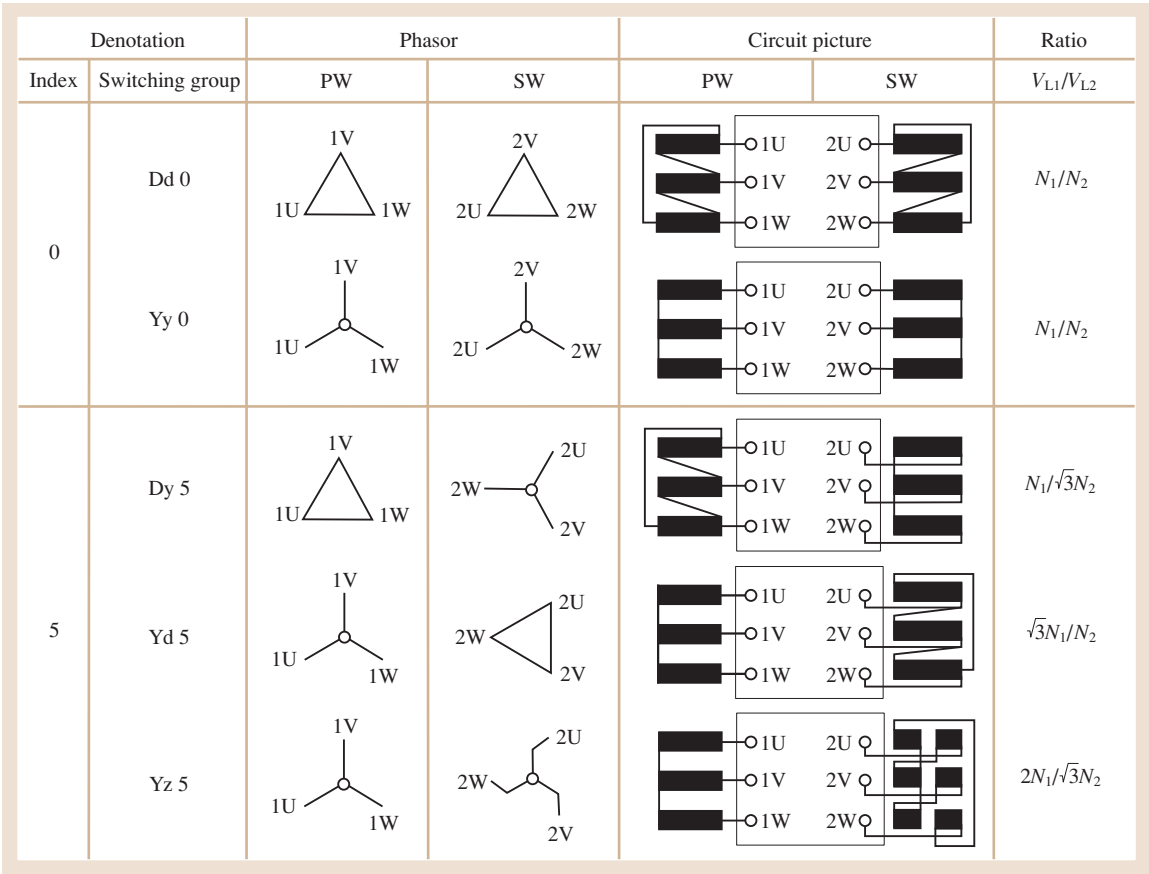


Fig. 17.32 Connection groups of three-phase transformers [17.1]

vantage is the direct electric connection between the high- and low-voltage sides.

17.2.2 Instrument Transformers

The instrument transformer is a device that uses the principles of a transformer to convert network currents and voltages into quantities that can be processed by other measurement instruments, protective or metering devices or control systems (Fig. 17.29). The norm for voltage measurement is 100 V, whereas for current measurement it is either 1 or 5 A. The secondary side of the instrument transformer is galvanically separated and, in particular, this property makes it possible to measure high-voltage systems. The total measurement error of these devices depends on the sum of the magnitude error and angle error. Depending on the accuracy levels of the instrument, they are grouped into several classes which characterize the allowable percentage error (class 0.1, 0.2, or 1.0).

Current Transformers

The primary and secondary windings in a current transformer are coupled by a ferromagnetic sheeting or toroidal core with low leakage inductivity. If high currents are measured, the core with the secondary winding is mounted around the primary winding. This way the primary winding usually consists of one turn. The resistance (burden) connected to the secondary side is used for the current measurement.

Since the accuracy is strongly depending on the existing of the current I_0, metal sheeting with high permeability in the operating range is used. Additional errors occur, when the core gets saturated by primary currents containing DC components.

Voltage Transformers

A voltage transformer works basically in open-circuit mode of the secondary winding. Because of this, it can be treated as an ideal transformer in which a simple connection between the measured input and read output appears. This relation can be described with the help of a basic expression which results from working in an open-circuit connection.

17.2.3 Three-Phase Transformers

Three-phase transformers consist of the three primary and three secondary windings coupled on one core (Fig. 17.31) [17.10]. Advantages of three-phase transformers are that they cost less, weigh less, and require less floor space in comparison to three separate single-phase transformers. The core of a three-phase transformer consists mainly of magnetic steel sheeting. In order to avoid magnetization losses silicated steel sheets with a thickness of 0.35 mm are mainly used.

For three-phase transformers the two elementary core forms, three and five limb, are applied. The windings consist of insulated copper wires. The power transformers are located in the transformer shell and are immersed in oil, which simultaneously maintains the insulation and cooling of the windings. Very often air can also be used as a cooling agent.

Transformer windings are classified in terms of their groups of connections (Fig. 17.32) [17.9]. The high-voltage side is written in upper-case letters and the low-voltage side is written in lower-case letters and for both groups an integer number is used to describe the angular difference in terms of 30° multiples between the phasors of the low- and high-voltage sides of the three-phase transformer.

The nominal short-circuit voltage for a three-phase transformer is defined in the same way as a single-phase transformer. The primary voltage excites the nominal current in the primary windings at short circuit on the secondary side (see Fig. 17.26).

The voltage difference between the primary and the secondary sides results from the appearing resistance, the leakage inductance, and the character of the load

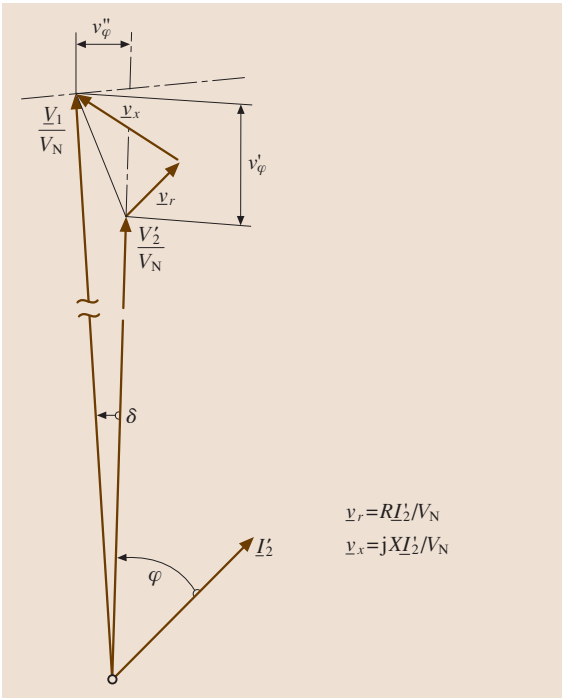


Fig. 17.33 Vector diagram for voltage change calculations

on the secondary side of the transformer. At nominal current, the relative voltage change with respect to the nominal voltage is defined as (Fig. 17.33)

$$\Delta v = v'_\varphi + 1 - \sqrt{1 - v''^2_\varphi}\,, \tag{17.138}$$

where

$$\begin{aligned} v'_\varphi &= v_\mathrm{r}\cos\varphi + v_\mathrm{x}\sin\varphi\,, \\ v''_\varphi &= v_\mathrm{x}\cos\varphi - v_\mathrm{r}\sin\varphi\,. \end{aligned} \tag{17.139}$$

The quantities v_r and v_x are the relative ohmic and inductive voltage drops, respectively.

The electrical data for transformers up to 40 MVA are standardized. By tapping the windings, the output voltage level can be regulated. This is used to ensure the rated voltage level during long-time changes of the load connected to the transformer.

17.3 Rotating Electrical Machines

17.3.1 General Information

Electrical machines convert mechanical energy to electrical energy (generator) or electrical energy to mechanical energy (motor) [17.11]. Energy conversion in almost all motors and generators takes place when a change in magnetic flux is associated with mechanical motion.

Each machine consists of at least one stationary part, the stator, and one rotating part, the rotor. According to Faraday's law, voltages are generated in windings or coils by:

1. Rotating a magnetic field past the windings or coils
2. Rotating the coils through a magnetic field, or
3. By designing the magnetic circuit so its reluctance changes with rotation of the rotor

Machine Types

In general, electrical machines can be divided into three types based on their mode of operation [17.1].

Induction Machines. As a general rule induction machines consist of a three-phase stator winding fed with AC currents, and a short-circuit three-phase rotor winding. Power is transmitted from the rotating magnetic field generated by the stator windings, through the air gap, to the rotor, which rotates asynchronously. A special form of this machine may be provided with slip-rings to gain access to the rotor windings.

Synchronous Machines. The majority of synchronous machines have a stator three-phase winding, and a rotor winding energized by a DC source that rotates at a constant speed proportional to the ratio of the applied frequency and the number of poles. For medium-sized and large machines a field coil is used on the rotor, while for small machines permanent magnets may be used instead.

Direct-Current Machines. A direct current (DC) machine may be viewed as a synchronous machine inside out. The commutator winding is placed at the rotor whereas the magnetic flux is produced in the stator. This can be achieved with a field coil or permanent magnets. The AC voltages induced in the rotor windings are converted to DC voltages through the action of the mechanical commutator. The commutator is mounted on the rotor shaft and connects groups of coils to the external terminals by means of brushes.

Nonrotating Machines. Linear machines are nonrotating machines, and their construction can be either induction or synchronous. They are used as short or long stator linear motors.

Types of Construction and Shaft Heights

The different types of construction for rotating electrical machines are described in IEC 60034.

Machines for industrial use, especially three-phase asynchronous (induction) machines, are produced according to IEC 60072. The construction of the machine is connected with nominal power and protection degree.

Degrees of Protection

Rotating machines are classified according to their size and protection provided by the enclosure as follows:

- Protected against human contact with parts that are under voltage
- Protected against contact and ingress of foreign bodies
- Protected against water ingress

This protection is described in IEC 60034-5. The degree of protection is represented by a symbol formed from a number and a letter, which consists of: IP + two digits + sometimes extra letters, e.g., IP 23 S.

The first digit of the code is for protection against contact and ingress of foreign parts whose size ranges between 1 mm and 50 mm. The second is for protection against water ingress. The digit 0 means that this machine is not protected (in the order of 0–6 for the first digit and 0–8 for the second digit). The second digit indicates that the machine is protected against either dripping water, dripping water at a tangential deviation up to 15°, spray water, shower water, stream water or that the machine is protected from dipping or immersion.

The special letters in the IP code are:

- W – weatherproof machine
- S – machines that are checked for water protection while stationary
- M – machines that are checked for water protection during operation

Losses and Efficiency

Using IEC 60034-2:1972 (IEC 60034-2A:1974 + A1:1995 + A2:1996) the total losses in an electrical

machine are determined as the sum of the following individual losses:

- Exciter circuit losses (only for DC and synchronous machines)
- Constant losses (iron, friction, and airing losses)
- Load-dependent losses (current heat losses)
- Load-dependent losses (other)

Let P_L be the sum of all losses. The efficiency of a machine η is defined as the ratio between the power output P_2 and total input power P_1

$$\eta = \frac{P_2}{P_1} = \frac{P_2}{P_2 + P_L} = \frac{P_1 - P_L}{P_1} . \qquad (17.140)$$

Efficiency as a function of the load is characterized by the maximum value η_{max}; this is the working point at which the load-independent losses are equal to the load-dependent losses. Machines for normal use are designed in such a way that η_{max} occurs for a load value below the nominal load, e.g., $P_2 = (7/8)P_N$.

Heating and Cooling

The heating of a machine (especially of the windings) has to be limited in order to guarantee the lifecycle of the machine. For this reason temperature limits exist, which depend on the class of insulation material used. The allowable temperature increase is given at an ambient air temperature of 40 °C. The hotspot temperature is considered based on the measurable temperature. The temperature limits rise for machines with insulation classes E, B, F, and H are specified in IEC 34-1. The given values are valid for windings in nominal operation and are specified using the resistance method. This method enables the engineer to determine the heat produced from increasing resistance and to calculate the temperature coefficient of the conducting material. The heat will be delivered to the primary cooling medium, in which it can be replaced or recooled using a secondary coolant. The cooling medium may be:

- Gaseous (air or hydrogen)
- Liquid (water or oil)

Duty Cycles

The heating of the machine is the result of magnitude and time-load characteristic. Four types of operating conditions are defined (Fig. 17.34):

- Continuous operation
- Short-time rating operation
- Periodic operation
- Nonperiodic operation

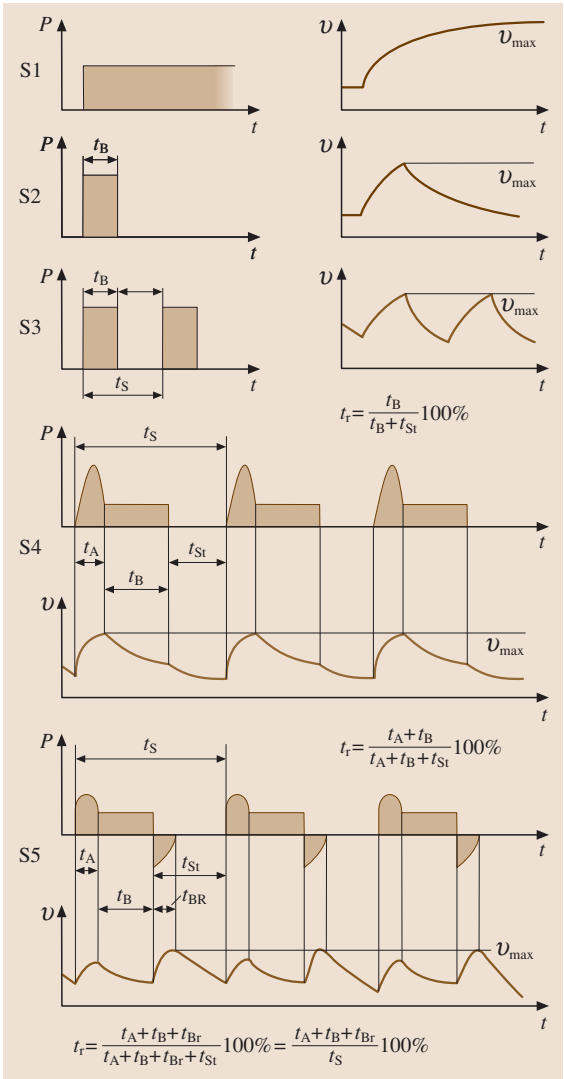


Fig. 17.34 Types of operation (S1 – continuous operation, S2 – short-circuit operation, S3 – impulse operation, S4 – impulse operation with influence of starting process, S5 – impulse operation with electrical acceleration) [17.1]

As one may expect, the same machine can be used more efficiently, compared to mode S1, when they are run in modes S2 or S3 because the upper temperature limit is not exceeded.

Vibrations and Noise

In electric machines mechanical oscillations occur as a result of magnetic excitation and imbalances. These oscillations can be assessed using ISO 2373.

The measured, effective oscillation speed v_{eff} also defines the peak value of the equivalent oscillation speed v_{eq} by taking into account the harmonic oscillations

$$v_{\text{eq}} = \sqrt{2} v_{\text{eff}} \,. \tag{17.141}$$

This value can be estimated using three degrees of oscillation strength:

- N – normal
- R – reduced
- S – special

As the name suggests, N is the most common oscillation strength for electrical machines. Accordingly, for machines with a size between $H = 132$ to 225 the critical value of allowable oscillation strength is $v_{\text{eff}} = 2.8\,\text{mm/s}$. Using the oscillating frequency f, the equivalent amplitude can also be calculated.

$$\hat{s} = \sqrt{2}/(2\pi f) v_{\text{eff}} \,, \tag{17.142}$$

where f is the frequency of the oscillation.

Noise in electrical machines can come from the following sources:

- Aerodynamic noise
- Magnetic noise
- Bearing and brushing noise

For environmental protection it is very important to construct low-noise electrical machines.

The noise perceived by human ears is dependent on the frequency and the amplitude of the noise source. It can be estimated by comparing the measured values with a weighting curve which gives the limit values of the noise (EN 60034-9). To verify the noise level, the sound pressure level L_{p} needs to be measured at idle speed so that the sound power level L_{W} can be calculated.

$$L_{\text{W}} = L_{\text{P}} + 10\,\text{lg}(S/S_0) \,, \tag{17.143}$$

$$S_0 = 1\,\text{m}^2 \,, \tag{17.144}$$

where S is the envelope area.

Electromagnetic Interference

Due to commutation processes electrical machines produce long-term disturbances. Switching operation evokes noncontinuous disturbances. The requirements for disturbances are described in EN 55014. Therein the limit values of interference voltage in the frequency range 0.15–30 MHz and interfering power, or alternatively the interfering field strength, in the range 30–300 MHz are stated. Figure 17.35 presents the common-mode interference voltage and interfering power. Generally, disturbance problems belong to the area of electromagnetic compatibility (EMC) [17.12]. Machines are described as EMC compatible when they can function in an electromagnetic field without influencing the function and environment of other machines that work in the same field.

Rotating Fields in Three-Phase Machines

Synchronous and induction machines with a three-phase system are called three-phase or rotating-field machines. The stator carries three-phase windings, whose coil sides lie in slots.

The currents i_{a}, i_{b}, and i_{c}, which flow in the U, V, and W windings, respectively, make up the symmetrical

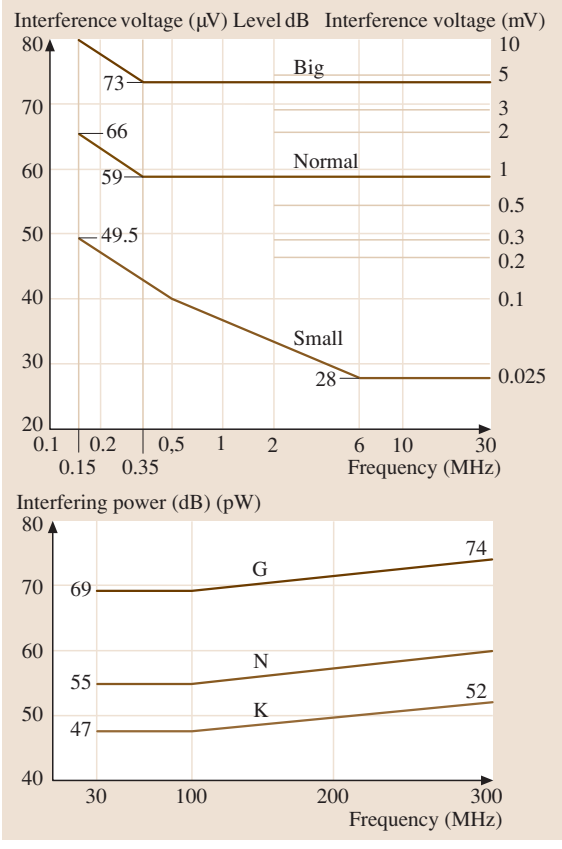


Fig. 17.35a,b Radio interference: **(a)** limit values of interference voltage; **(b)** limit values of interfering power by continuous interference [17.1]

three-phase system

$$i_a(t) = \hat{I}\cos(\omega t - \varphi)\,,$$
$$i_b(t) = \hat{I}\cos(\omega t - \varphi - 2\pi/3)\,,$$
$$i_c(t) = \hat{I}\cos(\omega t - \varphi - 2\pi/3)\,. \qquad (17.145)$$

These currents can also be described using a vector diagram, as shown in Fig. 17.36.

The currents produce the field excitation, which will be described through Ampère-turns. The Ampère-turns of three lines can be written

$$\Theta_{s,1}(\xi, t) = \Theta_{s,1}\cos(\omega t - \varphi - \xi)$$
$$\Theta_{s,1} = \frac{3}{2}\frac{4}{\pi}\frac{w\xi_1}{2p}\hat{I}\,, \qquad (17.146)$$

where w is the number of windings, ξ is the winding factor for the fundamental wave, and $2p$ is the number of poles in the machine.

This case can be also illustrated using the space vector method, as shown in Fig. 17.36. First, another complex surface has to be defined, which can be introduced as a sectional plane for a stator with two poles.

Next, the Ampère-current turns are defined

$$\underline{i}_s = i_a + i_b\,e^{-j\frac{2\pi}{3}}\,. \qquad (17.147)$$

Using the symmetrical system (17.147) $\underline{i}_s = \hat{I}_s\,e^{j\omega t}$ with $\underline{\hat{I}}_s = \frac{3}{2}\hat{I}\,e^{-j\varphi}$ and relating the phasor of the Ampère-turns

$$\underline{\theta}_s = \hat{\theta}_{s,1}\,e^{j\omega t} \quad \text{with} \quad \underline{\hat{\theta}}_{s,1} = \hat{\theta}_{s,1}\,e^{-j\varphi}$$

the location-dependent functional characteristic becomes, in connection with (17.146),

$$\theta_s(\xi, t) = \mathrm{Re}\left(\underline{\theta}_{s,1}\,e^{j(\omega t - \xi)}\right).$$

Figure 17.36 illustrates the current time phasor as well as the space vector of Ampère-turns. The space vector method is very useful for investigating the steady and dynamic operation of three-phase machines.

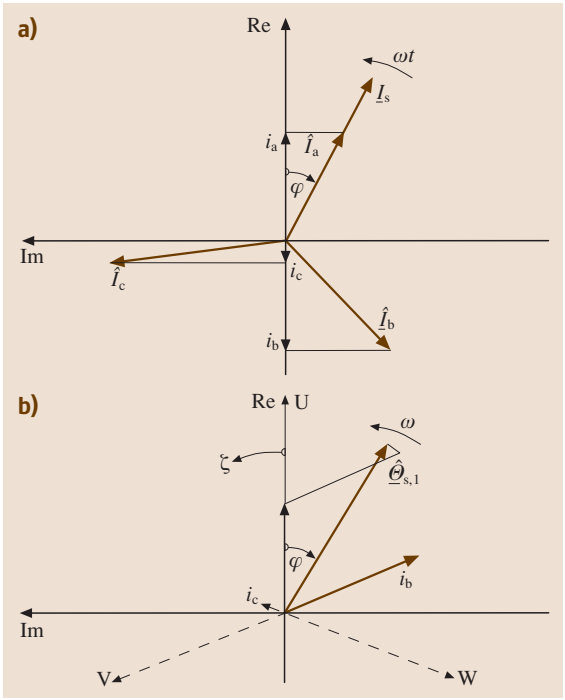


Fig. 17.36a,b Rotating field in rotating electrical machines. **(a)** Phasor diagram of currents; **(b)** space vector diagram of Ampère-turns

17.3.2 Induction Machines

Types

Induction machines are also called asynchronous machines because their operating speed is slightly lower than the synchronous speed in the motor mode and slightly greater than the synchronous speed in the generator mode.

Asynchronous machines (Fig. 17.37) are rugged, relative inexpensive, and require very little maintenance. They range in size from a few watts to about 10 000 hp. The speed of an induction motor is nearly constant, dropping by only a few percent when going from no load to full load. The significant disadvantages of induction machines are:

- The speed is not easily controlled
- The starting current may be five to eight times the full-load current
- The power factor is low and lagging when the machine is lightly loaded

The number of poles is usually 2, 4, or 6, and rarely 8 or 10.

The stator core and rotor windings of a three-phase induction machine are exactly like those of a synchronous machine. The only difference in construction between the two machines is in the rotor. In fact, an induction machine rotor of proper dimensions may be inserted in the bore of a synchronous machine stator, and the resulting induction machine will operate perfectly well.

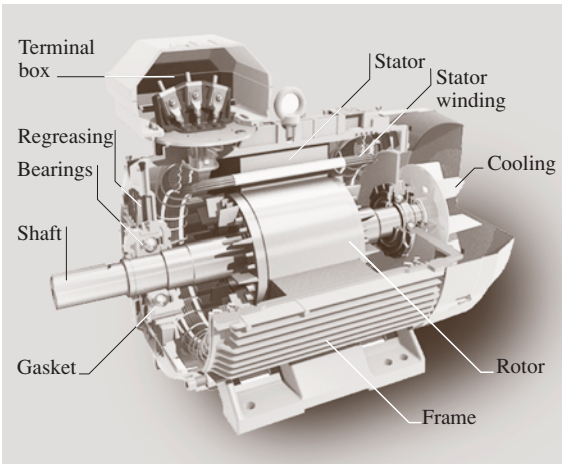


Fig. 17.37 Induction machine (ABB)

There are two types of induction machine rotors: wound and squirrel cage (short circuit). In either case, the rotor windings are contained in slots in a laminated iron core that is mounted on the shaft. In small machines, the rotor lamination stack is pressed directly onto the shaft. In larger machines, the core is mechanically connected to the shaft through a set of spokes called a *spider*.

The winding of a wound rotor is a polyphase winding consisting of coils placed in slots in the rotor core; it is almost always three phase and connected in a star format. The three terminal leads are brought to slip-rings mounted on the shaft. Carbon riding on these slip rings is shorted together for normal operation. External resistances are inserted into the rotor circuit to improve the starting characteristics. As the motor accelerates, the external resistances are gradually reduced to zero.

Squirrel-cage rotor windings consist of solid bars of conducting material in the rotor slots. These rotor bars are shorted together at the two ends of the rotor by end rings. In large machines, the rotor bars may be made of cooper alloy, driven into the slots and brazed into the end rings. The core laminations for such rotors are stacked in a mould, which is then filled with molten aluminum.

Equivalent Circuit Diagram and Circle Diagram

Figure 17.38 shown the equivalent circuit for an induction machine. The asynchronous machine with a squirrel-cage rotor is most suitable for such investigations, because the rotor resistance is independent of the slip. If such a machine is fed from a network with symmetrical voltage $\underline{V}_1$ and constant frequency f_1 the synchronous rotational speed n_s or angular speed Ω_s of the machine is defined as

$$\Omega_s = 2\pi f_1/p = \omega_1/p \tag{17.148}$$

or $n_s = 60 f_1/p$ in min^{-1}.

If this machine idles with asynchronous Ω then the rotor has the slip s opposite to the fundamental rotating field.

$$\begin{aligned} s &= 1 - (\Omega/\Omega_s)\,, \\ f_2 &= s f_1\,, \\ \omega_2 &= s\omega_1\,, \end{aligned} \tag{17.149}$$

where f_2 is the standardized (actual) frequency.

The equivalent circuit of the induction machine in Fig. 17.38 is similar to a transformer equivalent circuit.

The stator leakage inductance, magnetizing inductance, and rotor leakage inductance calculated on the stator side are described by the reactance parameters $X_{\sigma 1}$, X_n, and $X'_{\sigma 2}$ for the frequency ω_1. For the conversion of the rotor parameters to the stator side the factor ω_1/ω_2 is used. As a result the slip-dependent resistance R'_2/s appears on the rotor side. The current transfer locus (Fig. 17.39) describes the operating behavior.

The two significant points of this diagram are the idle speed point $P_0\,(s=0)$ and the point $P_\infty\,(s=\infty)$. Through a third point, e.g., the short-circuit point P_K at standstill of the machine ($s=1$) the circle is completely defined; its center is at point A and its diameter is given by the section $\overline{P_0 P_\Phi}$.

Similar to transformers, iron losses can be considered by using extra resistance in the cross branch in Fig. 17.38.

Operating Characteristics

The characteristic $M(\Omega)$ exhibits a breakdown torque M_k with the assigned breakdown slip s_k. The torque at $s=1$ is called the breakaway torque M_A. $M(s)$ can be

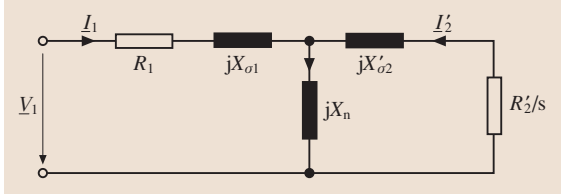


Fig. 17.38 Equivalent circuit of an asynchronous machine

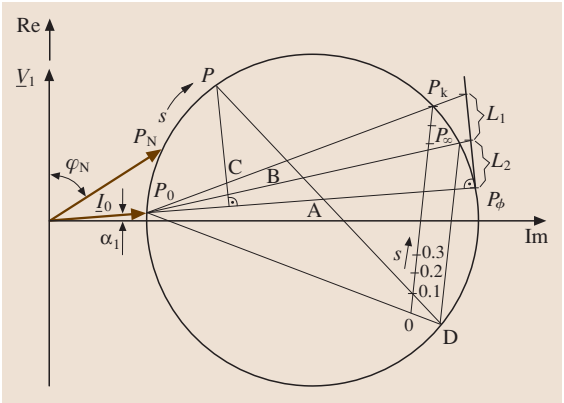


Fig. 17.39 Current transfer locus diagram

expressed as

$$M = M_k \frac{2(s/s_k)}{1+(s/s_k)} ,$$

with

$$s_k = \frac{R'}{X'_{\sigma 2} + X_{\sigma 1} \| X_n}$$

and

$$M_k = \frac{3}{2} \frac{V_1^2 (1-\sigma)}{\Omega_s \sigma X_1} \qquad (17.150)$$

by neglecting the stator resistance R_1.

According to IEC 60034, the relative breakdown torque M_k/M_N has to be greater than 1.6.

The machine works as a generator when the slip has negative values. At standstill the input current is equal to the short-circuit current, whose relative value I_A/I_N can be between 3 and 7 depending on machine construction.

The current of an idle running machine consists primarily of a reactive component which provides the required magnetization (Fig. 17.40). The operation characteristic (Fig. 17.40b) illustrates the interesting current, the power factor $\cos\varphi$, and the efficiency η depending on the mechanical output power.

The relationship between the input electric power P_{in} and the output mechanical power P_{out} of the induction motor is shown in the power-flow diagram in Fig. 17.41.

The input power to an induction motor P_{in} is in the form of three-phase voltages and currents. The first losses encountered in the machine are the Ohmic losses in the stator winding $3I^2R$ (the stator copper loss P_{SCL}).

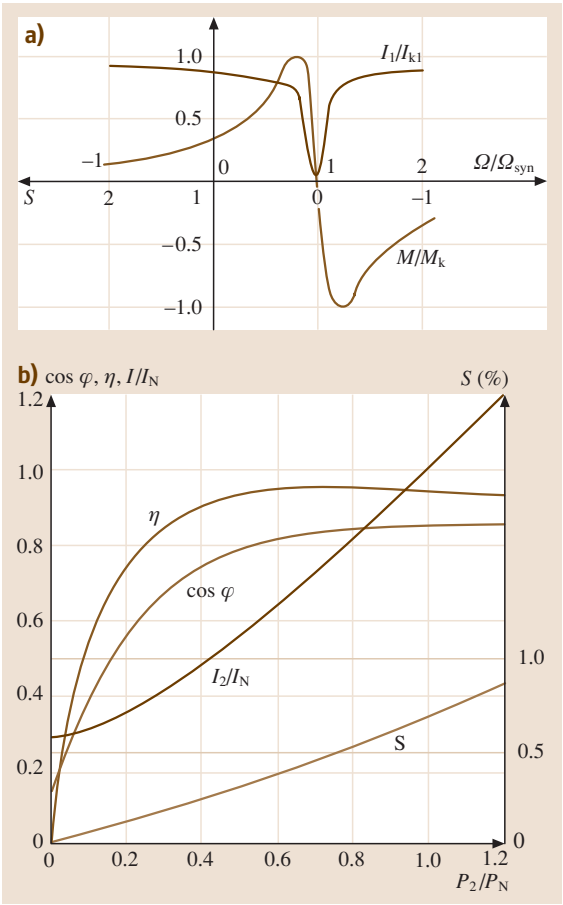


Fig. 17.40a,b Operation characteristic of an induction machine: **(a)** characteristic of current and torsion; **(b)** operation characteristic (load dependent) [17.1]

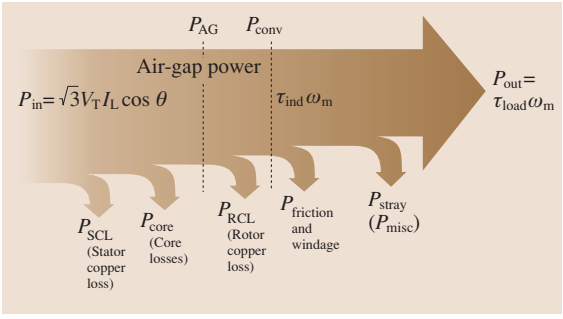


Fig. 17.41 Power flow diagram of an induction machine

The power lost as hysteresis and eddy currents in the stator is expressed as core losses (P_{core}). The power remaining at this point is transferred to the rotor across

the air gap between the stator and the rotor. This power is called the air-gap power P_{AG} of the machine. After the power is transferred to the rotor, some of it is lost as ohmic losses (the rotor copper loss P_{RCL}), and the rest is converted from electrical to mechanical power (P_{conv}). Finally, the friction and windage losses $P_{F\&W}$ and stray losses P_{misc} are subtracted. The remaining power is the output of the motor P_{out}.

Single-Phase Motors

In the examples above, it was assumed that the induction machines were fed with a three-phase symmetrical source.

Single-phase-fed induction motors, which are important in the small power sector, cannot produce a starting torque because of the missing rotating field in the stator.

To solve this problem, a split-phase motor is used in most cases. An extra start-up winding is arranged in addition to the major winding. Through a capacitor (capacitor start motor) or a higher resistance a phase shift of the current in the start-up winding is achieved. After reaching a defined speed the start-up winding can be switched of, e.g., with a centrifugal switch.

17.3.3 Synchronous Machines

Types

Like other electrical machines, synchronous machines may be operated either as a motor or as a generator.

In a synchronous generator, a DC current is applied to the rotor winding, which produces a rotor magnetic field. The rotor of the generator is then turned by a prime mover, producing a rotating magnetic field within the machine. This rotating magnetic field induces a three-phase set of voltages within the stator windings of the generator.

The rotor of a synchronous generator is a large electromagnet. The magnetic poles on the rotor can be:

- Salient – the magnetic pole sticks out of the rotor surface
- Nonsalient – the magnetic pole is constructed flush to the surface

Non-salient-pole rotors are normally used for two- and four-pole rotors. Because the rotor is subjected to changing magnetic fields, it is constructed of thin laminations to reduce eddy-current losses.

A DC current must be supplied to the field circuit on the rotor. Since the rotor is rotating, a special arrangement is required to provide the DC power to its field windings. There are two approaches to supply this DC power:

- From an external DC source to the rotor by means of slip rings and brushes
- From a special DC source mounted directly on the shaft of the synchronous generator

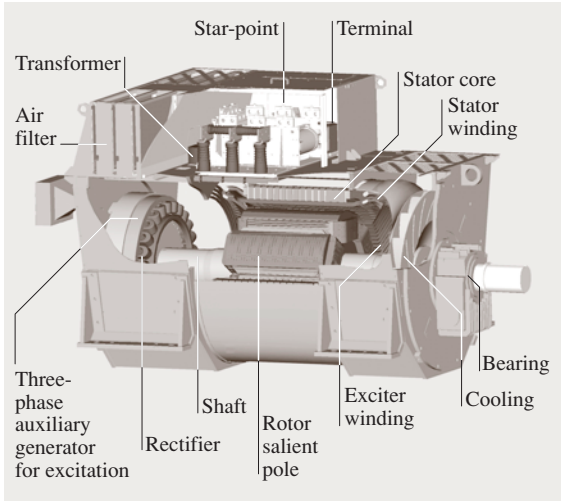


Fig. 17.42 Synchronous generator (ABB)

The largest electrical machines in the world are synchronous generators (Fig. 17.42). Some can produce as much power as 1700 MW. They are constructed as (two- or four-pole) turbogenerators and are driven, for example, by steam turbines. The power of a synchronous machine is limited by its possible rotor dimension (mechanical stress) and allowable armature current (temperature) (Table 17.3).

The limiting power values of two-pole turbogenerators at 50 Hz are described below.

Table 17.3 Synchronous generator power limits

Direct air cooling	80 MVA
Indirect air cooling	150 MVA
Hydrogen cooling without compressors	250 MVA
Hydrogen cooling with 5 bar overpressure	800 MVA
Water cooling, two pole	1200 MVA
Water cooling, four pole	1700 MVA

Operating Characteristics

In order to analyze the electrical behavior, d–q system modeling can be used. In this system it is assumed that the excitation axis is defined as the longitudinal axis (d) and the electrically orthogonal axis is the diagonal axis (q).

The polar wheel angle ϑ is typically used to characterize operation on a network at constant frequency. It represents the electrical angle between the phasor of the terminal voltage $\underline{V}_1$ and the phasor of the magnet wheel voltage $\underline{V}_p$. This magnet wheel voltage is the fictitious induced voltage that would result only from excitation without considering the armature reaction of the current $\underline{I}_1$.

The polar wheel angle ϑ is zero at idle speed. It has a positive value in generator operation and a negative value in motor operation.

Constant values of the voltage and excitation and sinusoidal torque $M(\vartheta)$ result in a circular transfer locus of the current (in Fig. 17.43, neglecting the resistance R_1). Thereafter the torque characteristic shows a breakover point, which depends on the polar wheel voltage in both generator and motor operation.

The quantity X_q/X_d is on the order of 0.7 for large generators and synchronous motors. Figure 17.44a shows the static stability limit of a synchronous generator, which restricts the steady operation area under excitation.

The V-curves can be used to allocate values of the stator current I_1 to the polar wheel voltage V_p (Fig. 17.44b).

Short-Circuit Characteristics

The short-circuit ratio of a synchronous generator is defined as the ratio of the field current required for the rated voltage at open circuit to the field current required for the rated armature current at short circuit. Although the short-circuit ratio adds no new information about the generator beyond what is already known from the saturated synchronous reactance, it is important to know what it is, since the term is occasionally encountered in industry.

In this section, we will survey the terminal three-phase short circuit in a machine with a dashpot cage. The output state is the idle speed at voltage V. Figure 17.45 shows the characteristic of the short-circuit current, which can be approximated by two exponential functions.

In this diagram the transient reactance X'_d and the subtransient reactance X''_d have to be defined, as well as the synchronous reactance, which determines the uninterrupted short-circuit current.

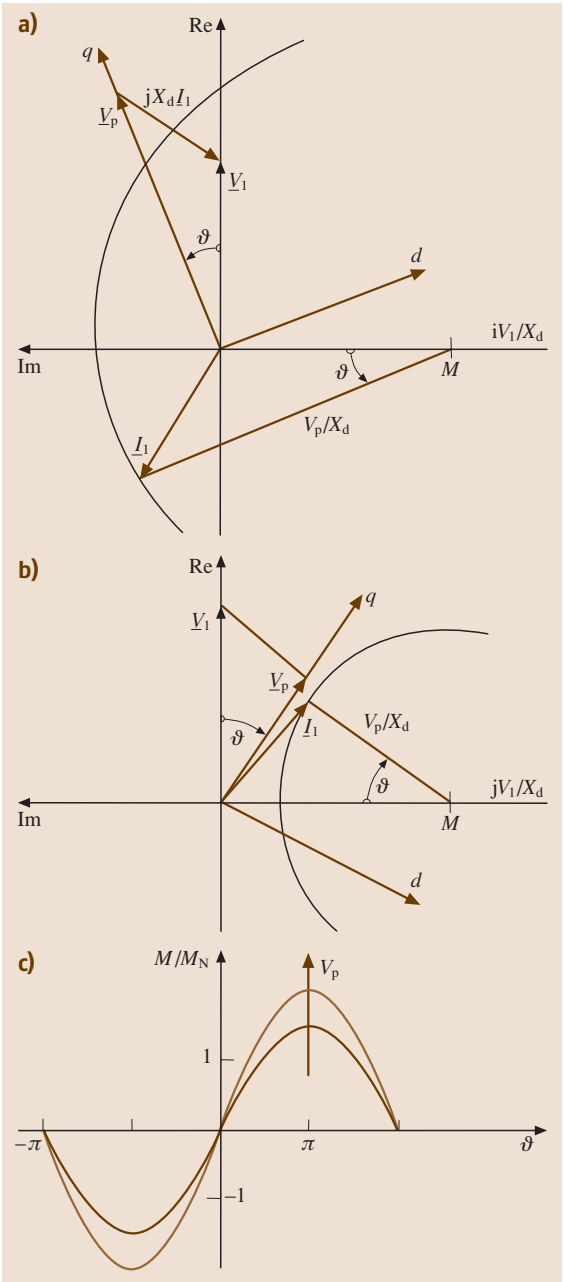


Fig. 17.43a–c Phasor diagram and torque characteristic of a synchronous machine. **(a)** Generator operation, overexcited; **(b)** motor operation, underexcited; **(c)** torque as a function of polar wheel angle

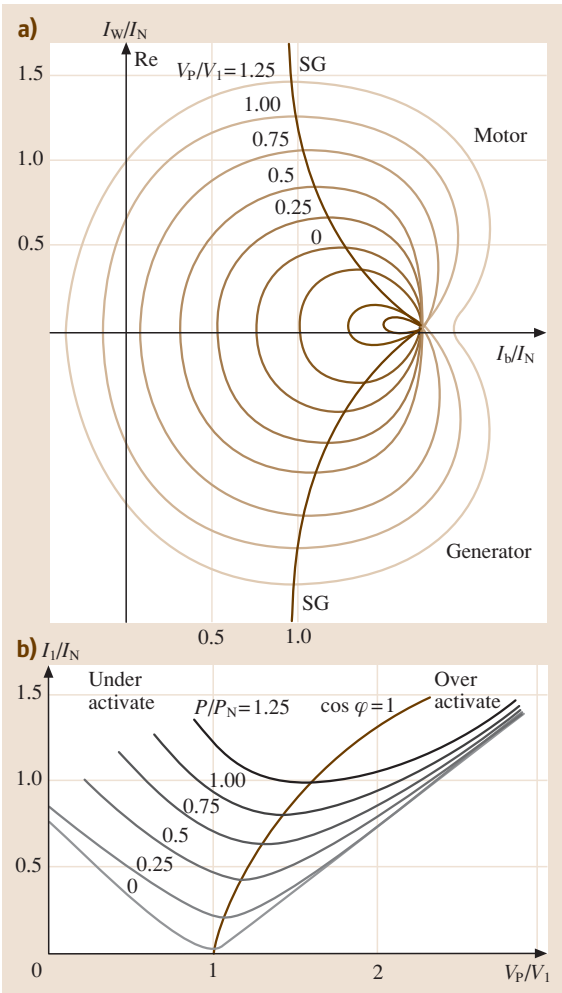


Fig. 17.44a,b Operation characteristic of an synchronous machine. **(a)** Current transfer locus diagram; **(b)** *V*-curves [17.1]

In this case the following parameters are important:

- the uninterrupted short-circuit current

$$I_k = V/X_d \,, \tag{17.151}$$

- The maximum asymmetric short-circuit alternating current transient

$$I'_k = V/X'_d \tag{17.152}$$

- The maximum asymmetric short-circuit alternating current

$$I''_k = V/X''_d \tag{17.153}$$

- The maximum asymmetric short-circuit current

$$I_P = \sqrt{2}\kappa I''_k \approx \sqrt{2}\cdot 1.8 I''_k \tag{17.154}$$

17.3.4 Direct-Current Machines

Types

DC machines are most often used as motors. These motors with permanent magnets are used as auxiliary drives in cars. For industrial uses these machines have a power of 100 kW and engine speed of up to 3000 min^{-1}. They are used as machine tools, hoisting devices, and machines in the basic material and paper industries.

Every DC machine has an armature winding *A* and, except for machines with permanent magnets, one excitation winding. This excitation winding can be implemented as a separate excitation winding *F*, as a shunt excitation *E*, or a series excitation *D* (Fig. 17.46). The armature current flows through the interpole winding *B*, which is responsible for the commutation process. Machines that place high demands on the dynamic system also have a compensation winding Fig. 17.47, which compensates the field of the armature. In this case a current slew rate of $(\text{d}i_A/\text{d}t)/I_N$ of up to 300 s^{-1} can be reached.

Machines for variable-speed drives are produced with laminated iron sheet to suppress the flux delays in the armature as well as in the stator.

Steady-State Operating Characteristics

The characteristic operation behavior of a DC machine is illustrated by the rotational speed and armature current as a function of torque. If constant losses are neglected the following relations apply

$$\Omega = V/c\Phi - R_A/(c\Phi)^2 M \,,$$
$$I_A = 1/(c\Phi)M \,. \tag{17.155}$$

Machines fed with constant voltage and, through shunt excitation, constant flux Φ, feature the typical shunt-wound operation characteristic.

The rotational speed is proportional to the feed voltage in unloaded operation and decreases linearly with increasing load due to the armature circuit resistance. At the same time, the armature current increases linearly, as shown in Fig. 17.48a. Otherwise, in a series-wound machines the flux is associated with the armature current through a curve that demonstrates a saturation characteristic. The unloaded rotational speed is only limited by frictional losses, whereas the rotational speed decreases rapidly with increasing load (Fig. 17.48b).

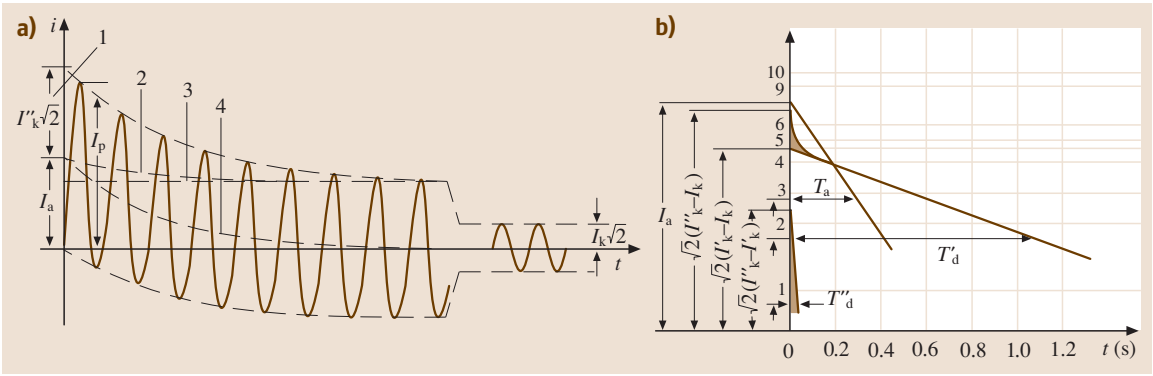


Fig. 17.45a,b Three-phase short-circuit characteristic: **(a)** current characteristic (1 – peak value of the surge short-circuit current, 2 – fast decay of the AC part, 3 – slow decay of the AC part, 4 – decay of the DC part); **(b)** interpretation of the short-circuit characteristic

Transient Operating Characteristics

The dynamic operation characteristics of DC machines play a very important role in industrial use where the dynamical demands are very high. In this section we will only consider machines with separate excitation ($\Phi = \text{const}$). In Fig. 17.49 a linear control circuit is shown. The input parameters are the imposed armature voltage v (the reference variable) and the load torque m_L (the disturbance variable).

The machine will be considered as a homogenous system with a total torque of inertia J. For this reason the structure of this system consists of a closed loop. This loop consists of an integrator with a mechanical time constant of T_M, connected in series to a first-order delay element with an (electrical) armature time constant of T_A. This second-order system has an electromechanical eigenfrequency w, where the periodic case with $d < 1$ results from the time constants

$$w = \omega_0\sqrt{1-d^2}\,,$$
$$\omega_0^2 = 1/(T_A T_M)\,,$$
$$d^2 = T_M/(4T_A) < 1\,. \qquad (17.156)$$

In this often occurring case the machine responds with damped oscillations when excited (e.g., by a step function). Figure 17.50 shows an example of the reference action at a standarised rotating speed of h_1 and

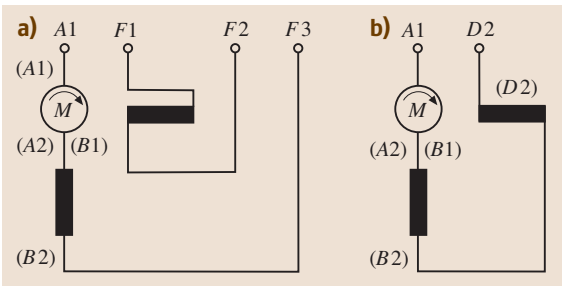


Fig. 17.46a,b Schema of a DC machine **(a)** with separate excitation (shunt wound); **(b)** with series excitation (series wound)

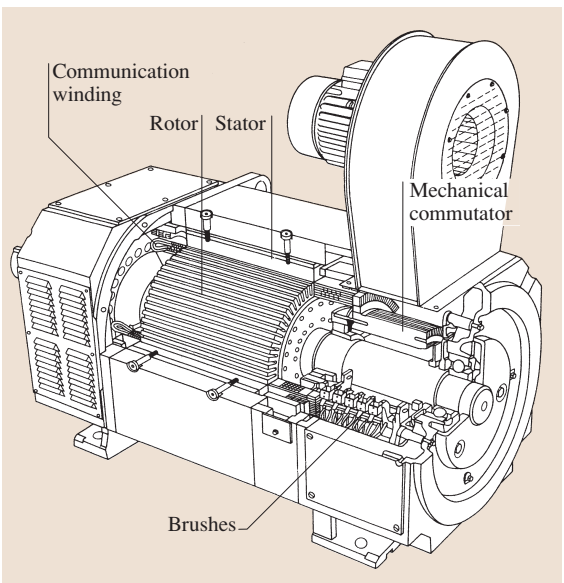


Fig. 17.47 DC motor (SIEMENS)

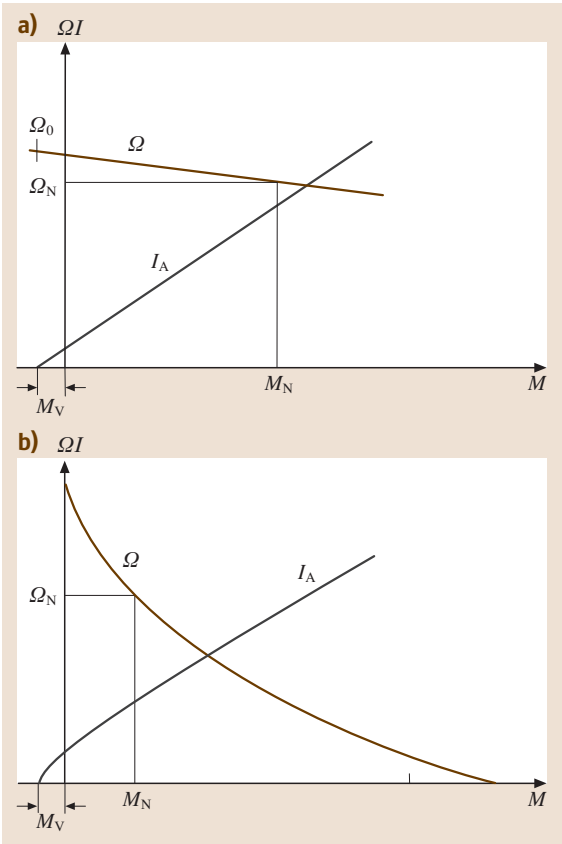


Fig. 17.48a,b Operation characteristic of a DC machine **(a)** with constant flux (shunt wound); **(b)** with series excitation (series wound)

armature current of h_2 after a step of the imposed armature voltage.

17.3.5 Fractional-Horsepower Motors

General Information

The power of fractional horsepower motors usually reaches 1 kW. These motors are used in large numbers in the consumer goods industry (household appliance, video and audio equipment, etc.), electrically driven tools, and as auxiliary drives in vehicles. Their professional use includes drives for data system technologies, office applications, and special drives for industry and scientific devices.

The construction of such motors mostly depends on their function, and not on standardized dimensions. The following types of motors are used in fractional-horsepower motors:

- Induction motors
- Synchronous motors
- Commutator machines

Small Induction Power Motors

In the following example asynchronous motors working with a one-phase system of 230 V and 50 Hz are considered. It is known that three-phase motors produce reduced torque when one phase is failing. However there is no start-up torque. External activation enables the motor to rotate in both directions. The mode of oper-

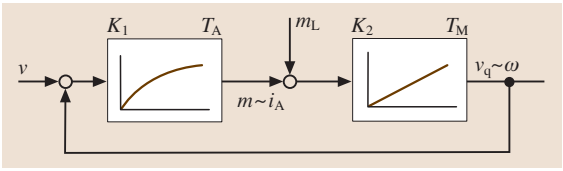


Fig. 17.49 Control circuit of a DC machine with constant flux. $K_1 = (C\Phi)/R_A$; $T_A = L_A/R_A$, $m = C\Phi i_A$; $v_q = C\Phi\omega$, $K_2 = R_A/(C\Phi)$; $T_M = J/(C\Phi)^2$

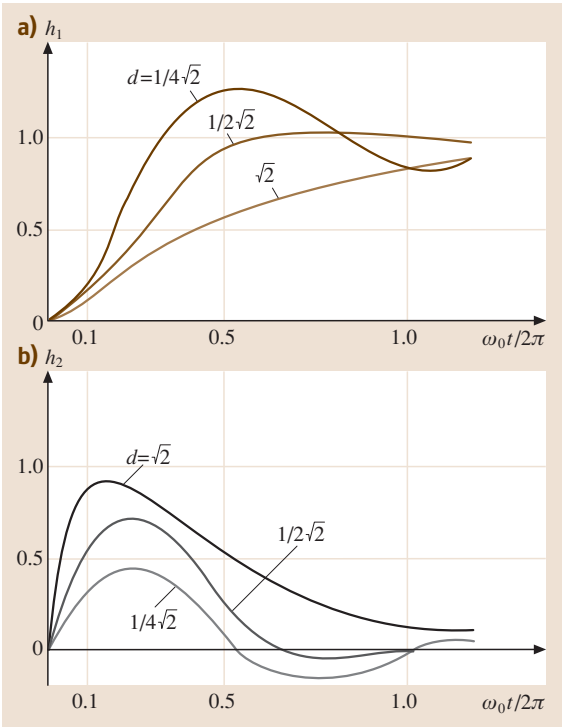


Fig. 17.50a,b Explanation of command action by step response: **(a)** rotation speed; **(b)** armature current [17.1]

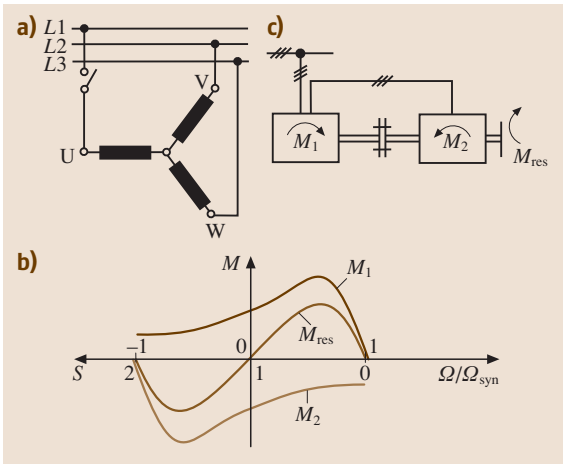


Fig. 17.51a–c One-phase operation of the machine because of phase breakdown. **(a)** Circuit; **(b)** characteristic of torsion; **(c)** equivalent system (two symmetrical machines for the negative- and positive-sequence system)

ation for a one-phase-fed motor can be explained using the symmetrical components theory (Fig. 17.10). The stator field features, in addition to the positive-sequence system, a negative-sequence system that works synchronously but in the opposite direction. Referring to the rotor the positive-sequence system rotates with a slip frequency of sf_1, whereas the negative-sequence system rotates with $(2-s)f_1$.

One-phase small power motors have a primary winding, a work winding, and an auxiliary winding. The current in the auxiliary winding produces an alternating field component that is shifted in position and time so that an incomplete rotating field is created. Figure 17.52 explains how the rotating field is produced, in which the positive-sequence system overbalances the negative-sequence system.

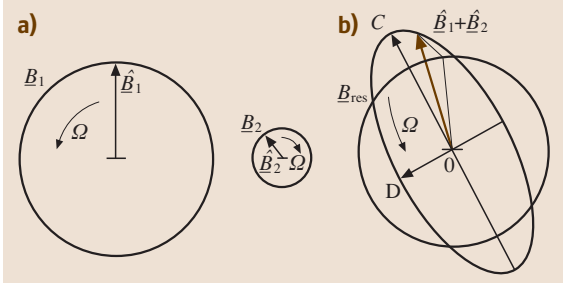


Fig. 17.52a,b Space vector for the creation of an elliptic rotational field. **(a)** Positive- and negative-sequence system; **(b)** superposition of both systems

Synchronous Small Power Motors for Network Operation

Three types of synchronous small power motors are in use:

- Permanent-magnet motors
- Hysteresis motors
- Reluctance motors

These motors can produce up to a few watts of power and can be used in switchgear equipment and me-

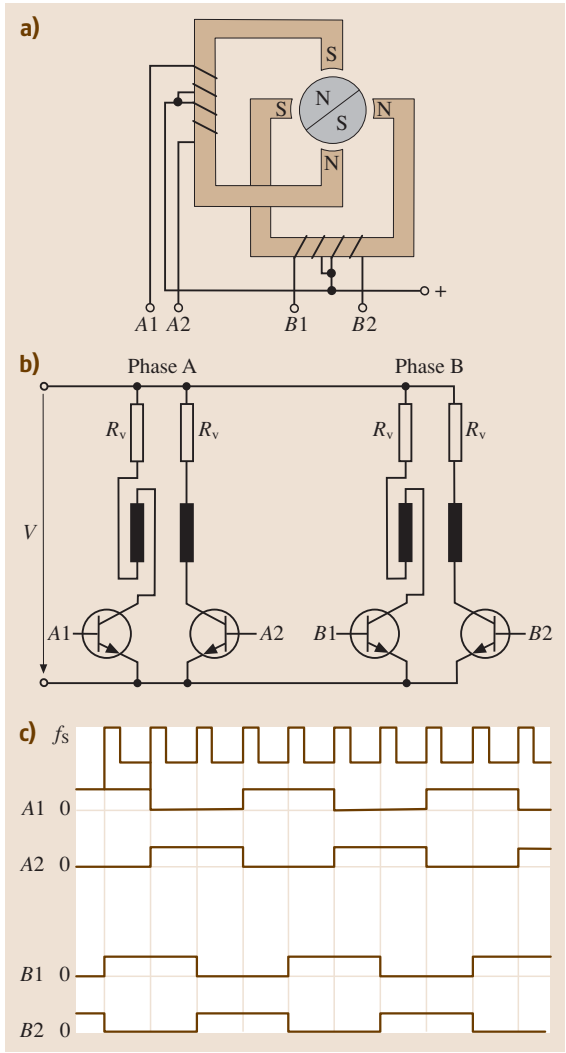


Fig. 17.53a–c Two-phase permanent-magnet stepper motor. **(a)** Principle of motor; **(b)** circuit for unipolar supply; **(c)** control in full step operation

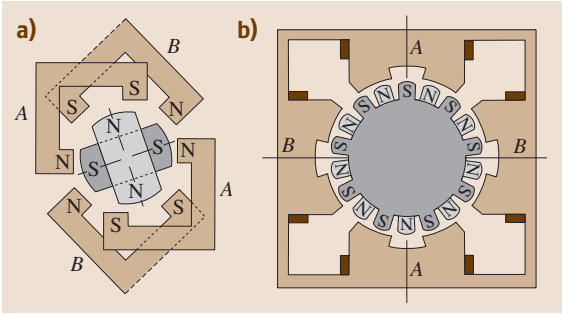


Fig. 17.54a,b Hybrid step motor (two phase). **(a)** Principle construction with two rotor plates ($z_r = 2$; $\alpha = 45°$); **(b)** construction with $z_r = 9$ and $\alpha = 10°$

chanical clocks. To produce a rotating (elliptical) field a capacitor-fed auxiliary phase as well as shaded pole windings can be used. For multipolar designs, claw pole constructions are favored.

Stepper Motors

Stepper motors may be thought of as polyphase synchronous motors fed with pulses produced by electronic controllers [17.13]. The name stepper comes from the feature that such a motor rotates only in defined steps (sectors of circle) for each pulse. Stepper motors can be excited permanently or by the principles of reluctance. In hybrid motors both principles are used to produce the torque. For large steps (e.g., 7.5−15°) claw pole machines are also applied. In order to realize very small steps (much smaller than 1°) polyphase hybrid motors are common. They are preferably equipped with high-grade magnets (e.g., cobalt-samarium).

The mode of operation can be explained using the example in Fig. 17.53, which depicts a two-phase stepper motor with a permanent-magnet rotor.

The drive is realized in a unipolar design through a four-transistor switch. At the default clock rate with the step frequency f_s the motor performs as shown in Fig. 17.53c. The hybrid motor contains an axial magnetized toric magnet in the rotor.

This magnet is placed between two magnetically soft rotor plates, which have z_r rims (Fig. 17.54). These plates are twisted against each other by half the width of the rims.

Electronic Commutated Motors

Generally speaking, these motors can be compared to synchronous machines. The rotor has a permanent excitation and the stator consists of multiple line windings and is controlled using an electronic controller. Unlike stepper motors, which are run in an open control chain, the drive for electronic commutation motors is dependent on the measured rotor position (e.g., using Hall sensors), as shown in Fig. 17.55. Usually the control of the rotational speed is realized as in DC current machines, which is why these motors are also called brushless DC motors.

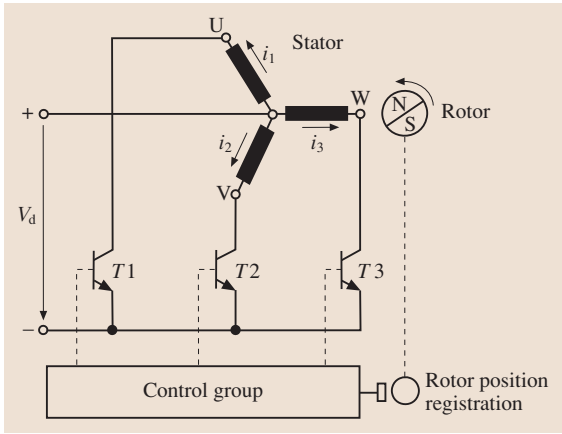


Fig. 17.55 Circuit of a brushless DC motor

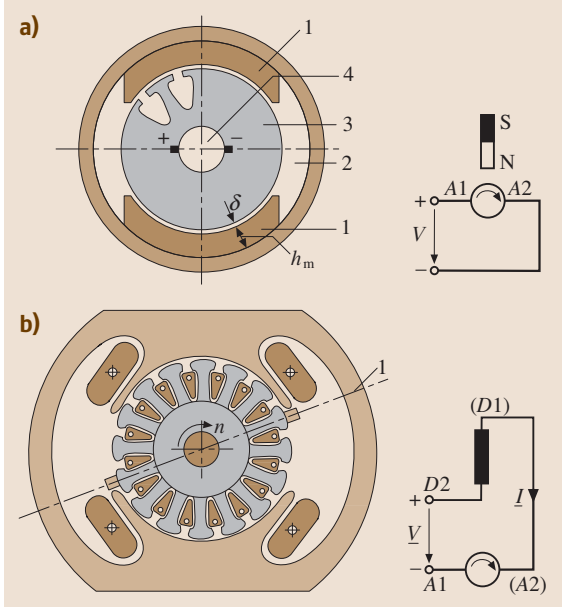


Fig. 17.56a,b Construction of **(a)** commutator small power motor ferromagnetic DC motor for 12 V: 1 – magnet part, 2 – iron interference, 3 – armature, 4 – commutator with brushes; **(b)** universal motor with movement of the brush axis: 1 – commutation axis

DC Small Power Motors

These kinds of motors are mostly used as auxiliary drives in vehicles and are supplied with a battery voltage of either 12 or 24 V. To produce low-cost motors only ferrite magnets are used. The high value of the temperature coefficient of the coercive field strength (about $+0.004\,\mathrm{K}^{-1}$) has to be considered as well as the deviation of the demagnetization characteristic in a high negative field strength; it has to be assured that at low ambient air temperature ($-20\,^{\circ}\mathrm{C}$) the start-up short-circuit current does not result in permanent demagnetization.

Universal Motors

These are motors which can work in both DC and AC systems. Nowadays they mostly use one-phase alternating current, especially in household appliances. The major advantage of these machines is that the rotating speed is not bound to the network frequency. For example, a vacuum cleaner that works at a rotating speed of up to $25\,000\,\mathrm{min}^{-1}$ can achieve a very high power-to-weight ratio. The universal motor is setup as shown in Fig. 17.56b.

The unloaded speed is limited by windage and friction. Because of their capability to operate at high speeds, universal motors of a given horsepower rating are significantly smaller than other kinds of AC motors operating at the same frequency.

Universal motors are ideal for home devices such as hand drills, hand grinders, food mixers, routers, and vacuum cleaners. Unfortunately the lifetime of such motors is limited by the lifetime of the brushes, which is about 2500 h.

17.4 Power Electronics

17.4.1 Basics of Power Electronics

Power Electronics is basically used to transform one electrical system into another. Major applications are variable-speed drives and power supplies. Configurations as follows are frequently met:

- Conversion DC → DC, e.g., to supply variable-speed DC drives from a constant-voltage DC line or as step-down DC–DC converters in decentralized power supplies
- Conversion AC → DC by a rectifier, e.g., to supply variable-speed DC drives or electronics – such as computers, communication equipment etc. – from AC mains
- Conversion DC → AC by an inverter, e.g., to feed energy generated from a DC source – such as a fuel cell or a photovoltaic generator – into AC mains
- Conversion AC → AC, e.g., to supply variable-speed AC motor drives by AC mains, or in generator mode to feed the generated electrical energy into mains

It is obvious that the voltage level and – in the case of AC – frequency are the subject of change in the transformation. The aforementioned conversion steps can of course also be cascaded, e.g., first rectifying mains AC into a DC system which is subsequently inverted again to supply a variable-speed AC drive.

Generally speaking, power electronics is used to transmit higher power than signal electronics, covering a range between some 100 W and several megawatts. Its operational principle can easily be derived based on following consideration, following the development of the control of variable-speed DC drives. A higher DC voltage u_1 can basically be reduced to a lower value $u_2 = R_1/(R_1+R_2)u_1 - R_1R_2/(R_1+R_2)i_2$ with a voltage divider circuit according to Fig. 17.57. It is however obvious that u_2 depends on the load current i_2; to keep the voltage u_2 constant for different loads, the voltage divider needs to be adapted. This can be achieved using controllable active devices instead of passive resistors. As an additional constraint power losses $p_V = (u_1-u_2)^2/R_1 + u_2^2/R_2$ in the circuit may become undesirably high and lead to low efficiency, especially in the case of a low ratio u_2/u_1. To avoid this, only ideally lossless passive devices, i. e., capacitors and inductors, should be used in power electronics along with digitally operated active devices.

Semiconductors are therefore used as switches according to Fig. 17.58a. They are either fully conductive, carrying a current $|i_S| \gg 0$ with negligible voltage drop $u_S \approx 0$, or fully blocking $|u_S| \gg 0$ with minimum leakage current $i_S \approx 0$; the transients between these steady states need to be short. Unless otherwise stated, the considerations of circuit theory in Sects. 17.4.2 and 17.4.3 therefore assume idealized behavior

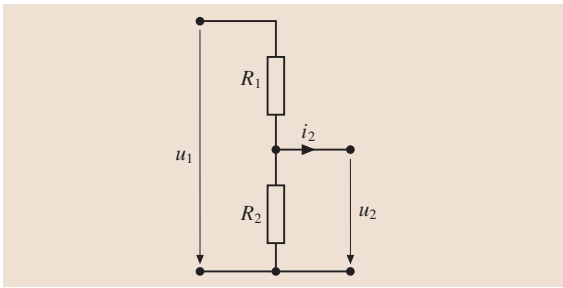


Fig. 17.57 Resistive voltage divider

of the basic power electronic switching components as follows:

- A insulated gate bipolar transistor (IGBT; Fig. 17.58b) carries a collector current $i_C \gg 0$ at ideally negligible saturation voltage $u_{CE} = u_{CEsat} \approx 0$ when turned on; when turned off, it will block a voltage $u_{CE} \gg 0$ at negligible leakage current $i_C \approx 0$. Control of the switching state is effected by the gate–emitter voltage u_{GE}.
- A metal oxide semiconductor field-effect transistor (MOSFET; Fig. 17.58c) behaves in a similar way. When turned off, it will block a voltage $u_{DS} \gg 0$ at negligible leakage current $i_D \approx 0$; it may however – like a diode, see below – conduct an inverse current $i_D \ll 0$. This also applies when turned on, where it will additionally conduct a drain current $i_D \gg 0$, both at ideally negligible voltage drop between the drain and source $u_{DS} \approx 0$. Control of the switching state is effected by the gate–source voltage u_{GS}.
- A diode (Fig. 17.58d) either carries a forward current $i_F \gg 0$ at an ideally negligible forward voltage $u_F = -u_R \approx 0$ or blocks a reverse voltage $u_R \gg 0$ at an ideally negligible reverse current $i_R = -i_F \approx 0$. Note that there is no separate control terminal: the switching state of a diode is thus simply defined by the mentioned voltage conditions in the power circuit.
- A thyristor (Fig. 17.58e) basically behaves similarly to a diode, however with additional means for control: after triggering with a gate current pulse $i_G > 0$ it can carry a forward current $i_T \gg 0$ at ideally negligible forward voltage $u_T = -u_R \approx 0$; this state is maintained until the forward current comes close to zero $i_T \to 0$ again, i. e., a thyristor device can be turned on via the gate control terminal but it will turn off only depending on the conditions in the power circuit. If not triggered, the thyristor will instead block a forward voltage $u_T = -u_R \gg 0$ at an ideally negligible forward leakage current $i_F \approx 0$ and will anyway block a reverse voltage $u_R \gg 0$, again at an ideally negligible reverse leakage current $i_R = -i_F \approx 0$ like a diode.

Transistors and diodes are relevant for the self-commutated circuits explained in Sect. 17.4.2, while diodes and thyristors are relevant for the externally commutated circuits described in Sect. 17.4.3. A more detailed description of these components with respect to their operation in these circuits is given in Sect. 17.4.4, which permits design issues for power electronics to be considered.

17.4.2 Basic Self-Commutated Circuits

DC Choppers

Stepping down a DC voltage u_Z to a lower value u_{1-} with $u_Z > u_{1-} > 0$ can be achieved with a buck chopper circuit as depicted in Fig. 17.59a operated according to Fig. 17.60a. When transistor T_1 is turned on and thus becomes conductive for $t_0 \le t < t_2$, the output voltage is $u_{L1-} = u_Z$. The voltage difference across the inductor

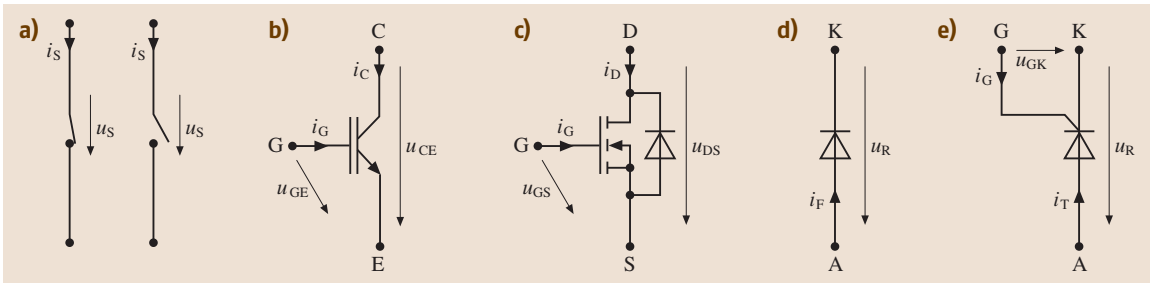


Fig. 17.58 **(a)** Switch turned on and off; **(b)** transistor (IGBT) with terminals C = collector, E = emitter, G = gate; **(c)** transistor (MOSFET) with terminals D = drain, S = source, G = gate; **(d)** diode with terminals A = anode, K = cathode; **(e)** thyristor with terminals A = anode, K = cathode, G = gate

L_{L1} determines the current change according to

$$L_{L1}\frac{di_{L1}}{dt} = u_{L1-} - u_{1-} \,. \tag{17.157}$$

Assuming that the DC link voltage and load voltage are constant $u_Z \equiv U_Z$, $u_{1-} \equiv U_{1-}$, the output current can be calculated to be

$$i_{L1}(t) = i_{L1}(t_0) + \frac{U_Z - U_{1-}}{L_{L1}}(t - t_0) \,, \tag{17.158}$$

which describes a linearly increasing waveform according to the aforementioned prerequisites. Because

$$i_{L1} = i_{S1} - i_{S2} \tag{17.159}$$

$i_{S1} = i_{L1}$ are identical while $i_{S2} = 0$ applies, because D_2 is the blocking reverse voltage.

After turning T_1 off at t_2, which leads to $i_{S1} = 0$, the inductor current can only be supplied via D_2 as $i_{S2} = -i_{L1}$. As long as this state lasts $u_{L1-} = 0$ applies, leading to an output current waveform starting at $i_{L1}(t_2)$ from (17.158) and changing according to (17.157)

$$i_{L1}(t) = i_{L1}(t_0) + \frac{U_Z}{L_{L1}}(t_2 - t_0) - \frac{U_{1-}}{L_{L1}}(t - t_0) \,. \tag{17.160}$$

The current will thus decrease linearly with a rate of change that depends on the load voltage U_{1-}. Because of the aforementioned device characteristics, with $i_{S1} \geq 0$ through transistor T_1 and $i_{S2} \leq 0$ through diode D_2, $i_{L1} \geq 0$ must apply. As long as $i_{L1} > 0$ flows continuously in the positive direction, the state described by (17.160) lasts until the cycle starts again at t_4, when T_1 is turned on again. If, on the other hand, $i_{L1} = 0$ becomes zero at $t_3 < t_4$, it will remain zero ($i_{L1} = 0$) with an output voltage determined by the load voltage $u_{L1-} = u_{1-}$ for $t_3 \leq t < t_4$ with both switches blocking; the current flow in this case is obviously discontinuous.

Both operational modes are depicted in Fig. 17.60a, calculated for the same DC link voltage U_Z; in the case of discontinuous operation U_{1-} has been set to a higher value than in the case of continuous operation, leading to a flatter increase of current according to (17.158) and a steeper decrease according to (17.160).

With respect to control, the symbols cycle duration $T_P = t_4 - t_0$ and switching frequency $f_T = 1/T_P$ are introduced together with the duty cycle

$$a = \frac{t_2 - t_0}{T_P} \,. \tag{17.161}$$

The latter represents the portion of the cycle during which T_1 is turned on; $0 \leq a \leq 1$ applies. The average output voltage can then be calculated to be

$$U_{L1-} = \frac{1}{T_P}\int_{t_0}^{t_4} u_{L1-}\,dt = aU_Z \,, \tag{17.162}$$

which is a linear transfer function of the variable average output voltage U_{L1-} versus the duty cycle a at constant DC link voltage U_Z. Considering the current, steady-state operation is achieved when $i_{L1}(t_0) = i_{L1}(t_4)$ applies: this is always the case in discontinuous operation with $i_{L1}(t_0) = i_{L1}(t_4) = 0$, while the condition for continuous operation can be derived from (17.160) to be $0 = (U_Z/L_{L1})(t_2 - t_0) - (U_{1-}/L_{L1})(t_4 - t_0)$, which is fulfilled for

$$a = \frac{U_{1-}}{U_Z} \,. \tag{17.163}$$

Considering (17.162) this means that the continuous current waveform in Fig. 17.60a does not change position when the average output voltage is adjusted to be equal to load voltage $U_{L1-} = U_{1-}$; the reason is that the average voltage across the inductor L_{L1} will then be zero.

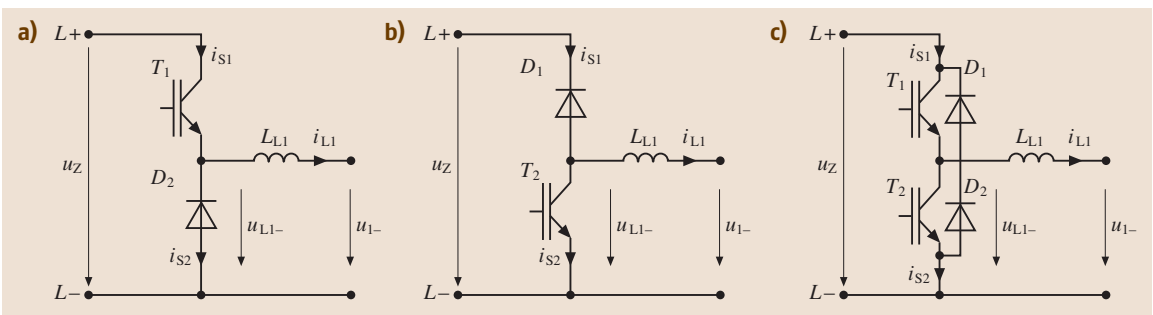


Fig. 17.59a–c Circuit diagrams of DC choppers with IGBTs and diodes: **(a)** buck chopper; **(b)** boost chopper; **(c)** phase leg

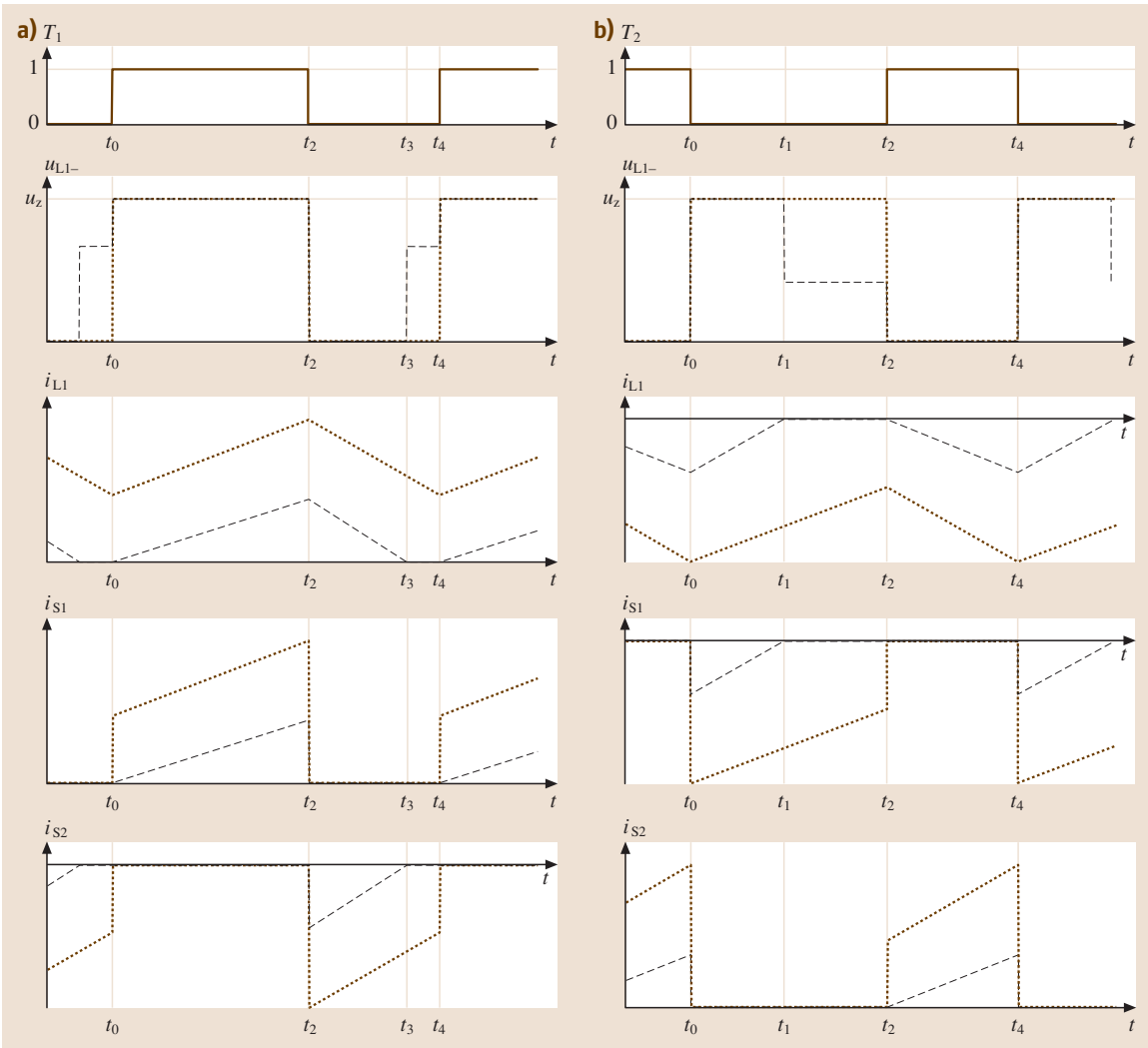


Fig. 17.60a,b Idealized waveforms of DC choppers in steady state – switching states of $T_1 - T_2$ (0 = off, 1 = on), output voltage u_{L1-} and current i_{L1}, switch currents $i_{S1} - i_{S2}$ of: **(a)** buck chopper and **(b)** boost chopper at continuous (*dotted*) and discontinuous (*dashed*) current flow

To transfer energy from a lower- to a higher-voltage source, the polarity of the current obviously needs to be inverted, which leads to the circuit diagram of the *boost chopper* in Fig. 17.59b with the corresponding waveforms in Fig. 17.60b. Here, D_1 permits a current flow $i_{S1} \leq 0$, T_1 $i_{S2} \geq 0$ and thus $i_{L1} = i_{S1} - i_{S2} \leq 0$.

To analyze the operation of this circuit it is assumed that T_2 will be turned on for $t_2 \leq t < t_4$. With $u_{L1-} = 0$, D_1 will then be blocking U_Z at $i_{S1} = 0$. Corresponding to the consideration of the buck chopper, (17.160) applies, i. e., $i_{L1} \leq 0$ will become more negative linearly in this time interval.

After T_2 has been turned off, leading to $i_{S2} = 0$, the output current $i_{L1} = -i_{S1} < 0$ will flow through D_1, making $u_{L1-} = u_Z$. As long as a current $i_{L1} < 0$ is flowing, (17.158) applies, meaning that i_{L1} becomes linearly more positive. Again corresponding to the above considerations, the current flow may be continuous, with $i_{L1} < 0$ not reaching zero for $t_0 \leq t < t_2$, or discontinuous with $i_{L1} < 0$ for $t_0 \leq t < t_1$ and $i_{L1} = 0$ for $t_1 \leq t < t_2$.

Due to the equivalent behavior of the boost and buck chopper, the derived control behavior – namely (17.162) and (17.163) – applies for both.

It is evident that combining the circuits of the buck and boost chopper according to Fig. 17.59a,b will permit bidirectional current flow and thus energy transfer from the higher- to the lower-voltage side and vice versa; the resulting *phase leg* circuit is depicted in Fig. 17.59c, and the corresponding waveforms are shown in Fig. 17.61. In the time interval $t_{10} \le t < t_{20}$ the circuit operates as a buck chopper, where the cycles $t_0 \le t < t_4$ according to Fig. 17.60a repeat with switching frequency f_T; the steady state with $i_{L1}(t_0) = i_{L1}(t_4)$ has been adjusted by choosing the duty cycle a according to (17.163). Correspondingly the DC–DC converter operates as a buck chopper in steady state for $t_{50} \le t < t_{60}$, as shown in Fig. 17.60b. In the time intervals $t_{20} \le t < t_{30}$ and $t_{40} \le t < t_{50}$ the average output voltage U_{L1-} is reduced according to (17.162) through a smaller duty cycle a; this is effected with shorter turn-on intervals of the transistor T_1 than in the steady state. The reduced average voltage U_{L1-} leads to a reduction of the average inductor current I_{L1}. The current is thus regulated by means of output voltage control through adjustment of the duty cycle a. Note that in the time interval $t_{30} \le t < t_{40}$ – where the steady state is adjusted again – the current plurally crosses the zero line. This is different to the operation of a pure buck or boost chopper as discussed before, where the current cannot change polarity and thus becomes intermittent. The reason is that the two transistors T_1 and T_2 of the phase leg are turned on alternately; thus paths for both polarities of current are open according to (17.159). Positive currents $i_{L1} > 0$ flow as $i_{S1} > 0$ or $i_{S2} < 0$ through T_1 or D_2, respectively, and negative currents $i_{L1} < 0$ as $i_{S2} > 0$ or $i_{S1} < 0$ through T_2 or D_1.

The considerations in this section can thus be summarized as follows. DC choppers are able to transfer energy from a higher- to a lower-voltage level $u_Z \ge u_{1-} \ge 0$, or vice versa. The average output voltage U_{L1-} is adjusted by the duty cycle a with a linear transfer function according to (17.161) and (17.162). During continuous current flow the average output current I_{L1} through the inductor will remain constant if (17.163) is met; for lower a it will decrease or become more negative, whereas for higher a it will increase or become more positive.

A typical application of such circuits is variable-speed drives, where a constant voltage U_Z is used to supply a DC machine with terminal voltage u_{1-}. With torque depending on current and speed on voltage, the drive is essentially in motor mode during buck chopper operation with $i_{L1} > 0$ and in generator mode during boost chopper operation with $i_{L1} < 0$; because $U_{1-} > 0$ these modes can be adjusted for one direction of rotation. To change to the other direction, the machine could be connected to L_1 and L_+ instead of L_-, which can be effected by using a relay in the case of low dynamic requirements or through a second phase leg as explained in the following.

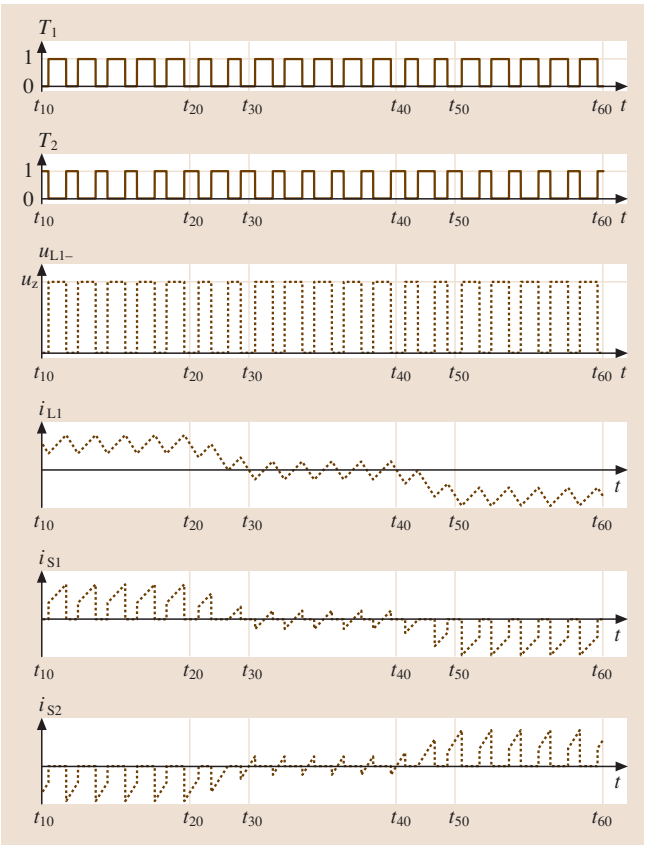


Fig. 17.61 Idealized waveforms of a two-quadrant DC chopper – phase leg: switching states of $T_1 - T_2$ (0 = off, 1 = on), output voltage u_{L1-} and current i_{L1}, switch currents $i_{S1} - i_{S2}$

Single- and Three-Phase Bridges

When combining two phase leg circuits according to Fig. 17.62a, the first consisting of T_1, D_1, T_2, and D_2 controlled by the duty cycle a_1 and the second consisting of T_3, D_3, T_4, and D_4 correspondingly by a_2, with

phase output voltages of

$$u_{\mathrm{L1M}} = \pm\frac{u_{\mathrm{Z}}}{2}\,,$$
$$u_{\mathrm{L2M}} = \pm\frac{u_{\mathrm{Z}}}{2}\,, \qquad (17.164)$$

which refer to a virtual center point M of the DC link, the resulting phase-to-phase output voltage will be

$$u_{\mathrm{L12}} = u_{\mathrm{L1M}} - u_{\mathrm{L2M}} = \begin{cases} \frac{u_{\mathrm{Z}}}{2} + \frac{u_{\mathrm{Z}}}{2} &= u_{\mathrm{Z}} \\ \frac{u_{\mathrm{Z}}}{2} - \frac{u_{\mathrm{Z}}}{2} &= 0 \\ -\frac{u_{\mathrm{Z}}}{2} - \frac{u_{\mathrm{Z}}}{2} &= -u_{\mathrm{Z}}\,. \end{cases} \qquad (17.165)$$

A calculation of the average values corresponding to (17.162) – assuming the same switching frequency f_{T} for both phase legs and a constant DC link voltage U_{Z} – yields

$$U_{\mathrm{L1M}} = \left(a_1 - \frac{1}{2}\right) U_{\mathrm{Z}}\,, \qquad (17.166)$$
$$U_{\mathrm{L2M}} = \left(a_2 - \frac{1}{2}\right) U_{\mathrm{Z}}\,. \qquad (17.167)$$

The difference is therefore

$$U_{\mathrm{L12}} = U_{\mathrm{L1M}} - U_{\mathrm{L2M}} = (a_1 - a_2)\, U_{\mathrm{Z}}\,, \qquad (17.168)$$

which covers the range $-U_{\mathrm{Z}} \le U_{\mathrm{L12}} \le U_{\mathrm{Z}}$ for the permissible values of a_1 and a_2 according to (17.161). This would fulfil the claim to provide both polarities of output voltage to supply a four-quadrant DC drive as stated at the end of last section.

The duty cycles a_1 and a_2 can be defined by several modulation schemes. A two-level modulation scheme is realized by switching the transistors in the diagonal of the bridge equally, i. e., T_1 and T_4 are on at the same time, leading to the first level of output voltage $u_{\mathrm{L12}} = u_{\mathrm{Z}}$, or T_2 and T_3 are on at the same time, leading to the second level $u_{\mathrm{L12}} = -u_{\mathrm{Z}}$, cf. the solid waveforms in Fig. 17.63a. The duty cycles are then $a_2 = 1 - a_1$, leading to an average output voltage of $U_{\mathrm{L12}} = (2a_1 - 1)\, U_{\mathrm{Z}}$ according to (17.168). In an alternative three-level modulation scheme independent duty cycles a_1 and a_2 can be adjusted, which makes all three levels of output voltage according to (17.165) occur, as depicted with dotted lines. Note that the average output voltage for the two different pulse patterns shown in Fig. 17.63a is the same: $U_{\mathrm{L12}} = -U_{\mathrm{Z}}/2$. The special case of a phase shift is illustrated in Fig. 17.63b. Both phase legs are operated with the same duty cycle, however the pulse patterns are shifted with respect to each other. Because $a_1 = a_2$, (17.168) yields an average output voltage of $U_{\mathrm{L12}} = 0$; it is thus a pure AC voltage. Its RMS value can be adjusted by using the phase shift, as is obvious from a comparison of the two displayed waveforms u_{L12}. The phase shift is commonly characterized by indicating the angle $2\pi\, f_{\mathrm{T}} t$. In

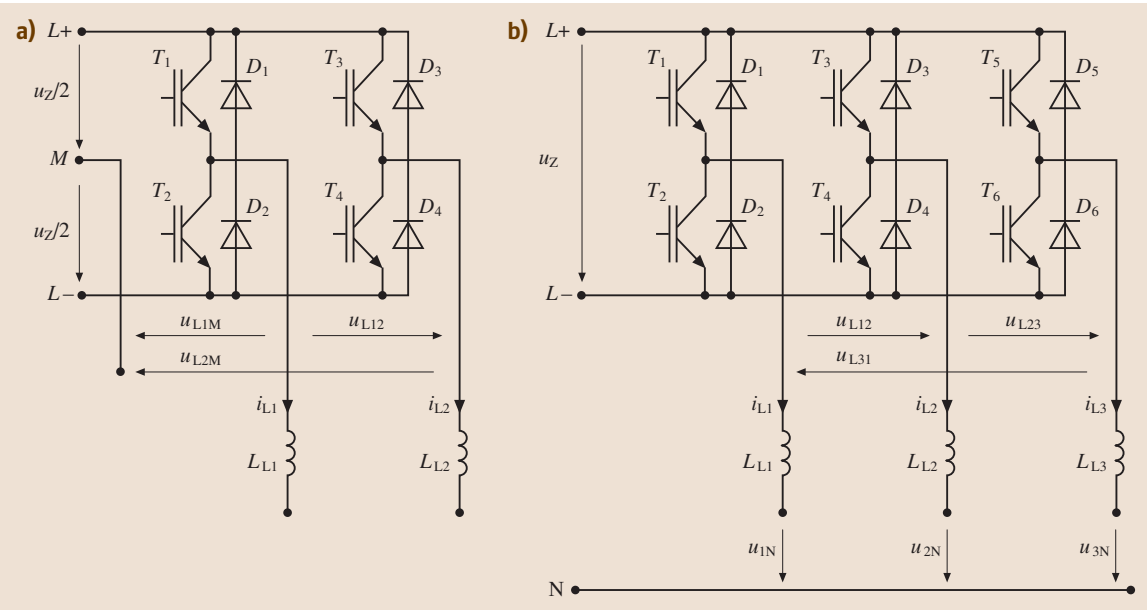


Fig. 17.62a,b Circuit diagrams of (a) single and (b) three-phase bridges with IGBTs and diodes

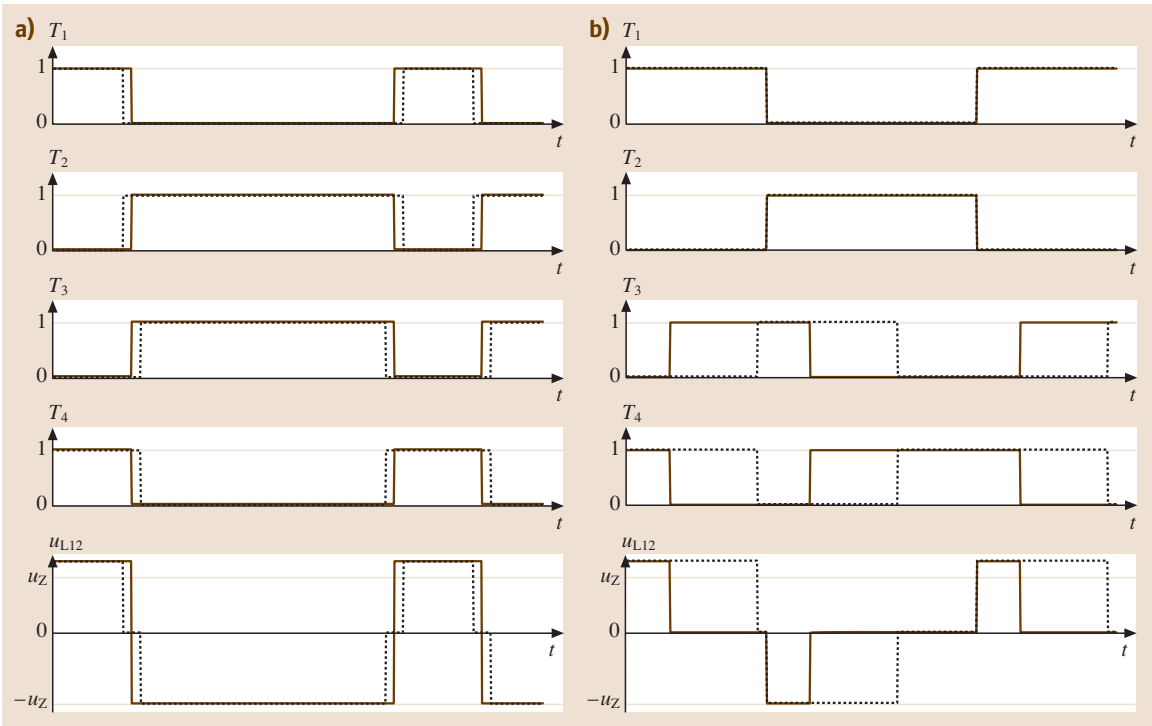


Fig. 17.63a,b Modulation schemes of a single-phase bridge – switching states of $T_1 - T_4$ (0 = off, 1 = on) and output voltage u_{L12}: **(a)** two-level (*solid*) and three-level (*dotted*) modulation scheme; **(b)** phase shift: 45° (*solid*) and 135° (*dotted*)

this way, the circuit can be used as a controllable source for a rectangular AC voltage, which is suitable, e.g., to feed a transformer. Consequently, the single-phase bridge under consideration is a key element of many switched-mode power supplies, where a DC voltage is inverted, then transformed to another voltage level with galvanic isolation and finally rectified.

A rectangular voltage shape can be divided up into a fundamental oscillation and harmonics; because of the latter however it is not well suited to the supply of electric machines with a rotating field or to feed electrical energy into an AC grid. Ideally, the converter would instead behave like a source of sinusoidal voltage. When the switching period of the converter is much shorter than the period of the desired output voltage $T_P \ll T_n$, this can be approximated by varying the duty cycle of the converter during each output cycle and in this way producing a sequence of average voltages over one switching period according to (17.168) which follows the sine wave on the output in a stepwise fashion. It is thus required to introduce pulse-width modulation [17.14]. This is frequently implemented by comparing a carrier signal u_T with switching frequency f_T to a control signal u_{St} with output frequency $1/T_n$, in this case a sine wave; if the control signal u_{St} is higher than the carrier u_T, one switching state of the respective phase leg or the bridge is applied, otherwise the other is applied. An example is shown in Fig. 17.64a, where a two-level modulation scheme has been implemented: it is obvious that the pulse lengths for which the transistors T_1 to T_4 are turned on vary during the displayed output cycle. In fact, the output voltage u_{L12} is mainly positive at the maximum of control signal u_{St}, mainly negative at its minimum, with a gradual change of the pulse widths in between; thus an average output voltage U_{L12} during one switching cycle is achieved. This emulation of a sine wave consequently leads to an output current i_{L1} with the typical ripple, but which basically follows a sine wave. Note that the output current i_{L1} depends on the voltage difference between the converter output voltage u_{L12} and the outer counter-voltage u_{12}, which in turn has been assumed to be sinusoidal. For one of the depicted counter-voltages, the converter will feed energy into the external AC source and thus oper-

ates in inverter mode; in the case of the other voltage, it takes energy from the source and thus operates as a rectifier. At a given external counter-voltage the current – including its phase angle with respect to the voltage, which determines the power factor $\cos\varphi$ – is determined by adjustment of the converter output voltage u_{L12}. It can therefore be concluded that the considered converter is not only suitable to transform one DC system into another but also to transform DC to AC as an inverter, and vice versa as a rectifier. It thus accomplishes several of the tasks outlined in Sect. 17.4.1. Single-phase bridges are frequently used as an interface between AC mains and a DC link: as inverters they may, e.g., feed electrical energy generated by renewable sources such as photovoltaic cells into the grid; as rectifiers, they may feed a DC link for power supplies. The harmonic content of the current – which needs to be limited in the case of mains connection [17.15] – can be controlled to be low.

Three-phase AC machines such as common induction or synchronous machines can be supplied with a corresponding three-phase converter circuit which is obtained when adding a third phase leg to the single-phase bridge, cf. Fig. 17.62b. Corresponding considerations then apply. Figure 17.64b exemplarily shows the generation of a three-phase system of output voltages u_{L12}, u_{L23}, and u_{L31}. Note that the odd number of phases requires the introduction of a three-level modulation scheme which may be implemented with one control voltage u_{St1}, u_{St2}, u_{St3} per phase. Three-phase converters of this kind are a key component of modern variable-speed drives: the machines can be supplied and thus controlled by three-phase AC systems with variable amplitude and frequency. Motor and generator operation of the machine are possible in inverter and rectifier mode of the converter, which is also able to supply the required reactive power. The direction of rotation is defined by the phase sequence of the converter output voltages, depending on the control voltages u_{St1}, u_{St2}, and u_{St3}. Consequently full four-quadrant operation of the drive can be achieved.

17.4.3 Basic Circuits with External Commutation

Uncontrolled Bridges

Although the circuits considered so far can be used as rectifiers, the conventional topology of rectifier bridges is often preferred. The reason is that it is simpler and needs no control, as is obvious from the schematics shown in Fig. 17.65.

Fig. 17.64a,b Idealized waveforms of pulse width modulated **(a)** single-phase bridge: from *top* to *bottom* – triangular carrier u_T (*solid*) and sinusoidal control signal u_{St} (*dotted*) – switching states of $T_1 - T_4$ (0 = off, 1 = on) – output voltage u_{L12} (*solid*) together with two different load voltages u_{12} (*dashed*) – output currents i_{L1} at these output voltages (*long dashed*: inverter operation, *short dashed*: rectifier operation); **(b)** three-phase bridge: from *top* to *bottom* – triangular carrier u_T (*solid*) and sinusoidal control signals $u_{St1/2/3}$ (*dotted and dashed*) – switching states of $T_1 - \overline{T}_6$ (noninverted 0 = off, 1 = on) – output voltages u_{L12}, u_{L23}, and u_{L31} ▶

To characterize their behavior when connected to the mains, mains voltages are defined to be

$$u_n = \hat{U}_n \sin(\omega t) \tag{17.169}$$

for the single-phase circuit or

$$u_{L1N} = \hat{U}_n \sin(\omega t)\,, \tag{17.170}$$

$$u_{L2N} = \hat{U}_n \sin\left(\omega t - \frac{2\pi}{3}\right)\,, \tag{17.171}$$

$$u_{L3N} = \hat{U}_n \sin\left(\omega t - \frac{4\pi}{3}\right)\,, \tag{17.172}$$

for the three-phase circuit, using the angular frequency of the mains $\omega = 2\pi/T_n$ with cycle duration T_n of one mains period. The RMS value of the single-phase or star voltages is therefore $U_{n,RMS} = \hat{U}_n/\sqrt{2}$. Unless otherwise stated, a constant DC current $i_d = I_d$ is further assumed.

Regarding the single-phase circuit it is obvious that in the upper half of the bridge the diode with the highest potential of anode, i. e., D_1 or D_3, will conduct i_d while the other is blocking $|u_n|$; correspondingly in the lower half of the bridge the diode with the lowest potential of cathode, i. e., D_2 or D_4, will conduct i_d while the other again blocks $|u_n|$. Consequently the switching states described in Table 17.4 occur. The corresponding waveforms are depicted in Fig. 17.66a: the output voltage u_{di} consists of two sinusoidal half waves $u_{di} = |u_n|$ per mains period T_n, which is why the circuit is called a two-pulse bridge connection, abbreviated to B2. Its average value can be calculated from (17.169) to be

$$U_{di0} = \int_{T_n} u_{di}(t)\mathrm{d}t = \frac{2}{\pi}\hat{U}_n = \frac{2\sqrt{2}}{\pi}U_{n,RMS}\,. \tag{17.173}$$

The input current i_n is rectangular. Note that a rectangular waveform can be expressed as the sum of

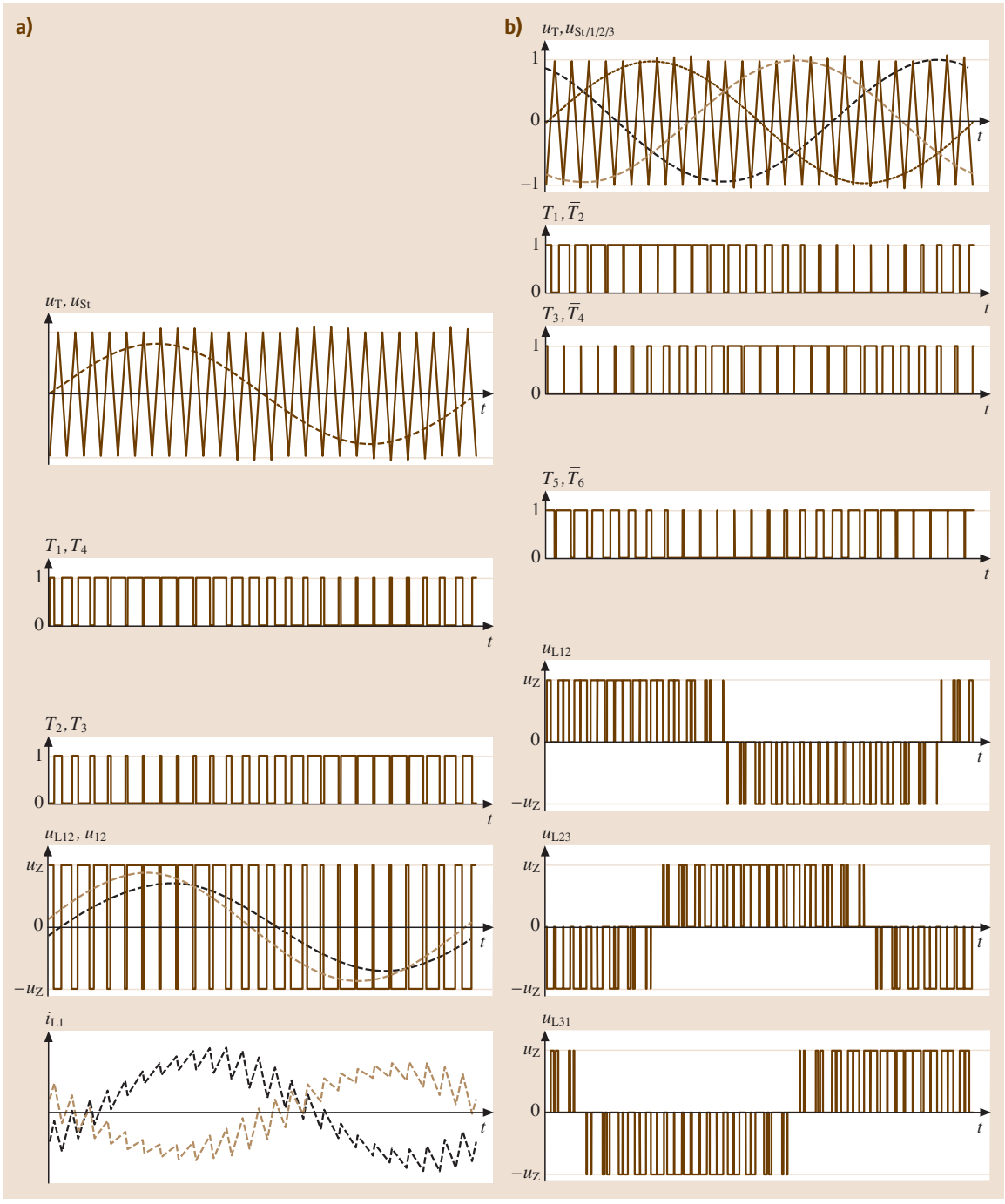


Table 17.4 States of the B2 bridge

Condition (uncontrolled B2)		
	$u_n \geq 0$	$u_n \leq 0$
Devices		
D_1	Conducts i_d	Blocks $-u_n$
D_2	Blocks u_n	Conducts i_d
D_3	Blocks u_n	Conducts i_d
D_4	Conducts i_d	Blocks $-u_n$
Input		
i_n	i_d	$-i_d$
Output		
u_{di}	u_n	$-u_n$

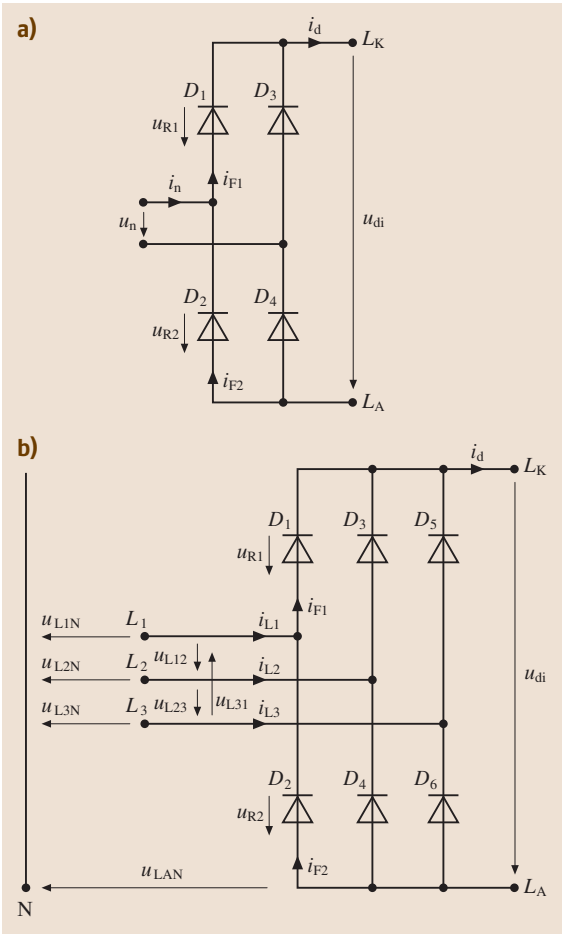


Fig. 17.65a,b Circuit diagrams of (a) single- and (b) three-phase uncontrolled rectifier bridges

a fundamental oscillation and harmonics; the occurrence of the latter in mains current is undesirable.

As stated, this basic consideration applies for the idealized circuit in Fig. 17.66a, comprising no inductances and being loaded with a constant DC current $i_d = I_d$. Frequently, the circuit is loaded by a buffer capacitor connected to the output terminals – which as a kind of voltage source might supply a self-commutated bridge as explained in Sect. 17.4.2, while inductors in the DC path are minimized. The shape of the current i_d then changes to pulses, as shown in Fig. 17.67b; their amplitude may be quite high compared to the load current, especially in the case of power-on when the previously discharged capacitor is charged. Note also that the shape of the DC voltage u_{di} is different, because all the diodes are blocking while $|u_n| < u_{di}$, which is buffered by the capacitor; the distribution of blocking voltages $u_{R1} + u_{R2} = u_{di}$ is not well defined by the circuit itself in this time interval.

The three-phase bridge shown in Fig. 17.65b behaves correspondingly to the described single-phase bridge. Again, the diode in the upper half of the bridge, i. e., D_1, D_3 or D_5, with the highest anode potential will be conducting, as will the diode in the lower half of the bridge, i. e., D_2, D_4 or D_6 with the lowest cathode potential, while the others are blocking. The six resulting states are summarized in Table 17.5 and illustrated in Fig. 17.67a. In this case the output waveform is composed of six pulses per mains period T_n, where the maximum of the noninverted and inverted line-to-line voltages is applied as the output voltage $u_{di} = \max\{u_{L12}, -u_{L12}, u_{L23}, -u_{L23}, u_{L31}, -u_{L31}\}$. Consequently the circuit is called a six-pulse bridge connection, abbreviated to B6. The average output voltage can be calculated to be

$$U_{di0} = \int_{T_n} u_{di}(t)\,dt = \frac{3\sqrt{3}}{\pi}\hat{U}_n = \frac{3\sqrt{6}}{\pi}U_{n,RMS}\,. \qquad (17.174)$$

Note that the line current waveforms of the B6 bridge consist of three levels including $i_{L1/2/3} = 0$.

To summarize it should be stated that the uncontrolled single- or three-phase diode bridges are only usable as rectifiers because the power $U_{di0} i_d \geq 0$ is exclusively transferred from the mains AC side to the DC side. The moment of current commutation from one diode to another is determined by the time values of the mains voltages. This is why the output voltage of the uncontrolled bridge – which is proportional to the mains voltage according to (17.173) or (17.174) – cannot be controlled. The inherently rectangular shape of the mains currents in the case of a constant DC load

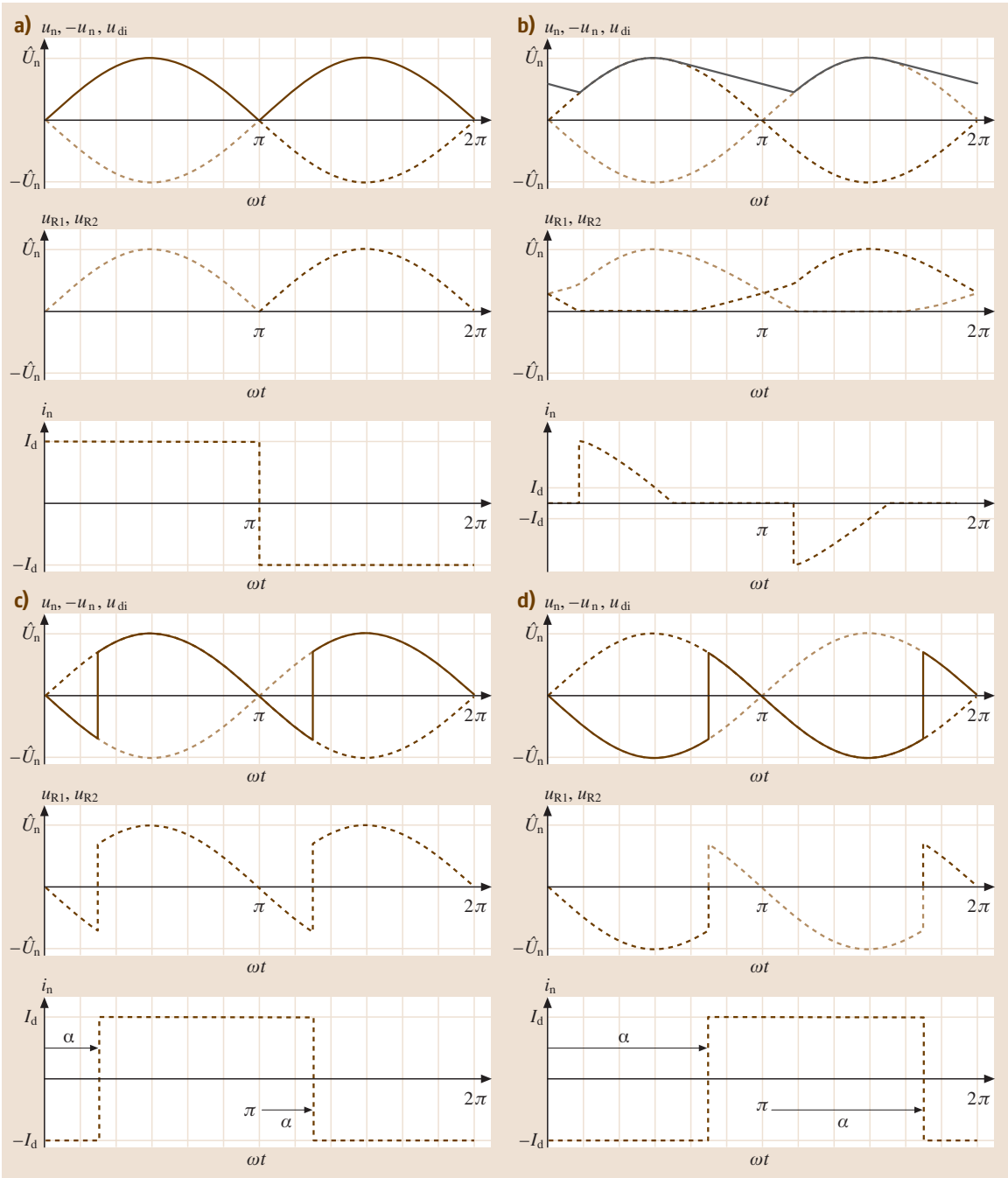


Fig. 17.66a–d Idealized waveforms of a B2 bridge: **(a)** uncontrolled or firing angle $\alpha = 0$ with constant DC current I_d; **(b)** uncontrolled with output capacitor in parallel with constant DC current I_d load; **(c)** controlled with $\alpha = \pi/4$ and constant DC current I_d; **(d)** controlled with $\alpha = 3\pi/4$ and constant DC current I_d: from *top* to *bottom* – line voltage u_n (*thick, dotted*), inverted line voltage $-u_n$ (*thin, dotted*), and output voltage u_{di} (*solid*) – reverse blocking voltages of diodes $D_1 - D_2$ or thyristors $T_1 - T_2$, respectively, u_{R1} (*thick*) and u_{R2} (*thin*) – line current i_n with label of firing angle if $\alpha > 0$

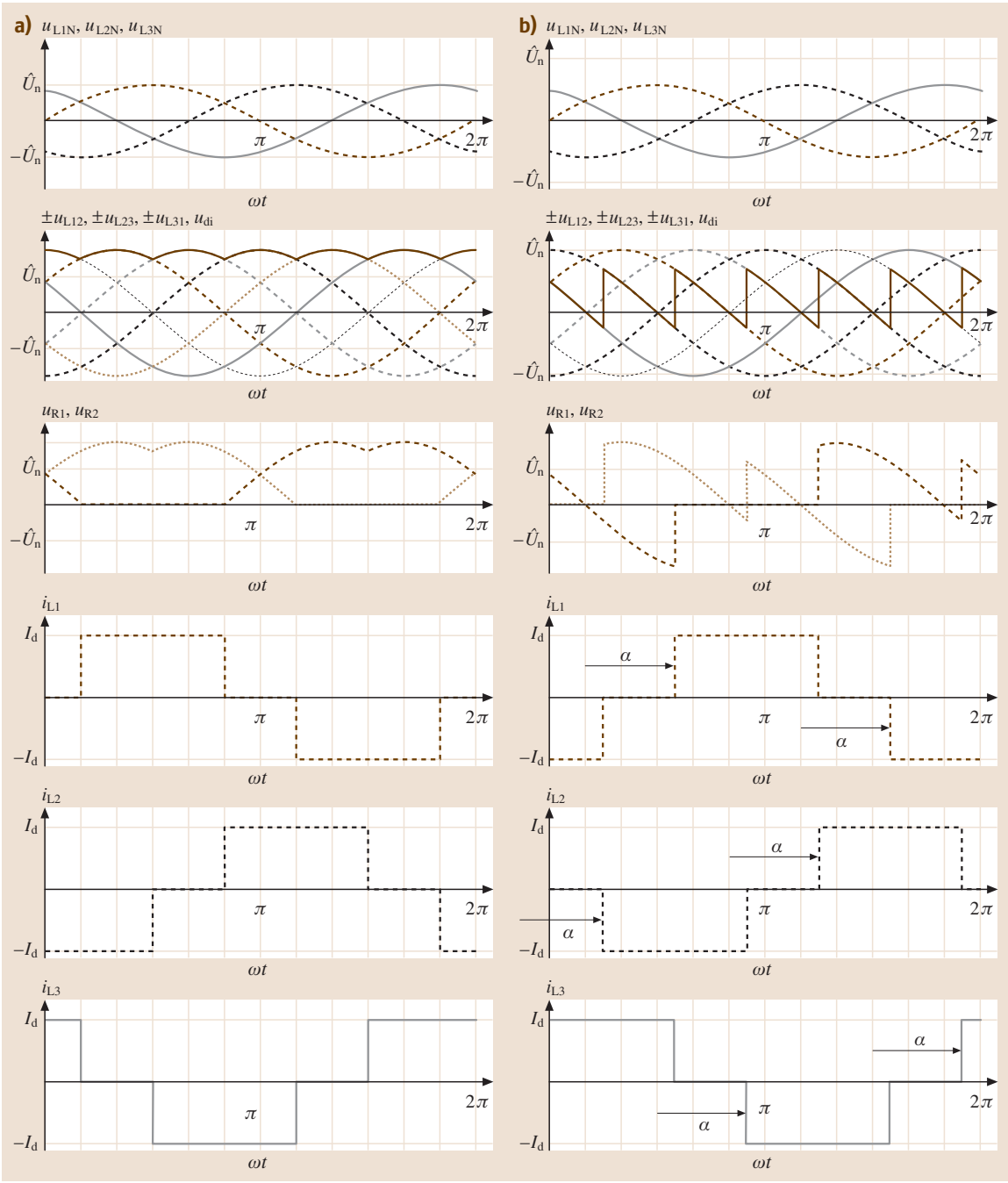


Fig. 17.67a,b Idealized waveforms of B6 bridge with constant DC current i_d: **(a)** uncontrolled or firing angle $\alpha = 0$; **(b)** controlled with $\alpha = 5\pi/12$: from *top* to *bottom* – three-phase star voltages u_{L1N}, u_{L2N}, u_{L3N} – three-phase line-to-line voltages u_{L12}, u_{L23}, and u_{L31} (*thick*, *dotted*, and *dashed*), inverted line-to-line voltages $-u_{L12}$, $-u_{L23}$, and $-u_{L31}$ (*thin*, *dotted*, and *dashed*) and output voltage u_{di} (*solid*) – reverse blocking voltages of diodes $D_1 - D_2$ or thyristors $T_1 - T_2$, respectively, u_{R1} (thick) and u_{R2} (*thin*) – three-phase line currents i_{L1}, i_{L2}, i_{L3} with label of firing angle if $\alpha > 0$

Table 17.5 States of the B6 bridge

Condition (uncontrolled B6)						
	$u_{L1N} \geq u_{L2N}$ $u_{L1N} \geq u_{L3N}$ $u_{L2N} \leq u_{L3N}$	$u_{L1N} \geq u_{L2N}$ $u_{L1N} \geq u_{L3N}$ $u_{L3N} \leq u_{L2N}$	$u_{L2N} \geq u_{L1N}$ $u_{L2N} \geq u_{L3N}$ $u_{L3N} \leq u_{L1N}$	$u_{L2N} \geq u_{L1N}$ $u_{L2N} \geq u_{L3N}$ $u_{L1N} \leq u_{L3N}$	$u_{L3N} \geq u_{L1N}$ $u_{L3N} \geq u_{L2N}$ $u_{L1N} \leq u_{L2N}$	$u_{L3N} \geq u_{L1N}$ $u_{L3N} \geq u_{L2N}$ $u_{L2N} \leq u_{L1N}$
Devices						
D_1	Conducts i_d	Conducts i_d	Blocks $-u_{L12}$	Blocks $-u_{L12}$	Blocks u_{L31}	Blocks u_{L31}
D_2	Blocks u_{L12}	Blocks $-u_{L31}$	Blocks $-u_{L31}$	Conducts i_d	Conducts i_d	Blocks u_{L12}
D_3	Blocks u_{L12}	Blocks u_{L12}	Conducts i_d	Conducts i_d	Blocks $-u_{L23}$	Blocks $-u_{L23}$
D_4	Conducts i_d	Blocks u_{L23}	Blocks u_{L23}	Blocks $-u_{L12}$	Blocks $-u_{L12}$	Conducts i_d
D_5	Blocks $-u_{L31}$	Blocks $-u_{L31}$	Blocks u_{L23}	Blocks u_{L23}	Conducts i_d	Conducts i_d
D_6	Blocks $-u_{L23}$	Conducts i_d	Conducts i_d	Blocks u_{L31}	Blocks u_{L31}	Blocks $-u_{L23}$
Input						
i_{L1}	i_d	i_d	0	$-i_d$	$-i_d$	0
i_{L2}	$-i_d$	0	i_d	i_d	0	$-i_d$
i_{L3}	0	$-i_d$	$-i_d$	0	i_d	i_d
Output						
u_{di}	u_{L12}	$-u_{L31}$	u_{L23}	$-u_{L12}$	u_{L31}	$-u_{L23}$

or the occurrence of pulse shapes is disadvantageous. Because of their simplicity and ruggedness these circuits are widely used for mains rectification, e.g., to supply drive converters or as input stages of power supplies.

Controlled Bridges

A means to control the output voltage is frequently desirable. Basically maintaining the previously discussed bridge circuits, this can be achieved by controlling the timing of current commutation through the use of thyristors, which behave similarly to diode but need to be triggered by a gate pulse (Fig. 17.58). The corresponding circuit diagrams are depicted in Fig. 17.68. The principle of operation of a single-phase thyristor bridge is illustrated in Fig. 17.66c. At $\omega t \to \pi$ the current will flow through T_1 and T_4, which is the state characterized in the middle column of Table 17.4; the input current is thus $i_n = i_d$ and the output voltage is $u_{di} = u_n$. In a diode bridge, the current would commutate to the other pair of diodes at $\omega t = \pi$; in a thyristor bridge this cannot happen before the respective thyristors are triggered. This means that the previous switching state is maintained until triggering is applied, at the firing angle $\omega t = \alpha$. The correspondent waveforms are shown in Fig. 17.66c for $\alpha = \pi/4$; the output voltage continues to follow the sine wave $u_{di} = u_n$ in its negative half wave. The average value thus becomes lower; depending on the firing angle α it can be calculated to be

$$U_{di\alpha} = \int_{T_n} u_{di}(t)\mathrm{d}t = \frac{2}{\pi}\hat{U}_n \cos\alpha$$
$$= \frac{2\sqrt{2}}{\pi} U_{n,RMS} \cos\alpha = U_{di0} \cos\alpha \,. \qquad (17.175)$$

Compared to the uncontrolled bridge the trigger delay by α thus leads to a reduction of the average bridge output voltage by a factor of $U_{di\alpha}/U_{di0} = \cos\alpha$. Note that, for $\alpha > \pi/2$, $U_{di\alpha}$ can in this way become negative; a corresponding example with $\alpha = 3\pi/4$ is depicted in Fig. 17.66d. Here, however, consideration of the thyristor blocking voltages – see u_{R1} and u_{R2} – becomes important: because $u_n < 0$, $u_{R1} = -u_n > 0$ will, for example, apply for $\pi < \alpha < 2\pi$ when T_3 is conducting, cf. also Table 17.4. As stated in Sect. 17.4.1, a thyristor with forward current flowing cannot be actively turned off to take over a forward blocking voltage, but it will block a reverse voltage $u_R > 0$ like a diode. It is thus necessary to trigger T_3 sufficiently long before $\alpha = \pi$ to guarantee that a positive reverse voltage $u_{R1} > 0$ will be applied to T_1; if this is not the case, T_1 and T_3 might conduct at the same time, leading to a potentially destructive short circuit, also called through conduction. If this is observed, the thyristor bridge will operate as an inverter, transferring power $U_{di\alpha} i_d < 0$ from the DC side to the mains for $\pi/2 < \alpha \ll \pi$, while it remains a rectifier transferring power $U_{di\alpha} i_d \geq 0$ from the mains to

the DC side for $0 \le \alpha \le \pi/2$. The shape of the current waveform is also rectangular, as analyzed previously; however it should be noted that the current waveform is shifted by the firing angle α as marked by the arrows in Fig. 17.66c and d. The occurrence of a phase angle between the mains voltage and current is equivalent to undesirable consumption of reactive power from the mains when voltage control is applied through $\alpha > 0$.

Again the three-phase circuit according to Fig. 17.68b behaves equivalently as depicted in Fig. 17.67b: thyristor triggering is delayed from the instant of natural commutation (where a diode would start to conduct) by the firing angle α. Consequently the output voltage follows the sine waves according to Table 17.5 in a delayed interval, again reducing the average value with respect to the uncontrolled bridge (17.174)

$$
\begin{aligned}
U_{\mathrm{di}\alpha} &= \int_{T_\mathrm{n}} u_\mathrm{di}(t)\mathrm{d}t = \frac{3\sqrt{3}}{\pi}\hat{U}_\mathrm{n}\cos\alpha \\
&= \frac{3\sqrt{6}}{\pi}U_{\mathrm{n,RMS}}\cos\alpha \\
&= U_{\mathrm{di}0}\cos\alpha\,. \qquad (17.176)
\end{aligned}
$$

The circuit can thus correspondingly be used as a controlled rectifier transferring power $U_{\mathrm{di}\alpha} i_\mathrm{d} \ge 0$ from the mains to the DC side for $0 \le \alpha \le \pi/2$, or as a controlled inverter, transferring power $U_{\mathrm{di}\alpha} i_\mathrm{d} < 0$ from the DC side to the mains for $\pi/2 < \alpha \ll \pi$. The above findings regarding the required safety margin $\alpha \ll \pi$ to avoid through conduction and regarding the shape and phase angle of the input current waveforms $i_{\mathrm{L}1}$, $i_{\mathrm{L}2}$, and $i_{\mathrm{L}3}$ apply for the three-phase bridge as well.

In conclusion, the considered thyristor bridges permit the control of the output voltage, which can be varied by the factor $\cos\alpha$ by using the firing angle $0 \le \alpha \ll \pi$. However, operation at $\alpha > 0$ leads to undesirable consumption of first-harmonic reactive power from mains, while the additional harmonic content of the input currents is the same as for uncontrolled bridges. This voltage control method permits the use of these circuits as rectifiers and inverters. They are thus suitable to supply DC machines from the mains; because the considered bridge is suitable only for one polarity of current, i. e., a two-quadrant drive, it must be further complemented to create two-way rectifiers enabling four-quadrant operation. Due to the replacement of DC by AC drives and the availability of self-commutated converters over a wide power range the use of fully controlled thyristor bridges has decreased. An alternative is half-controlled bridges, where half of the

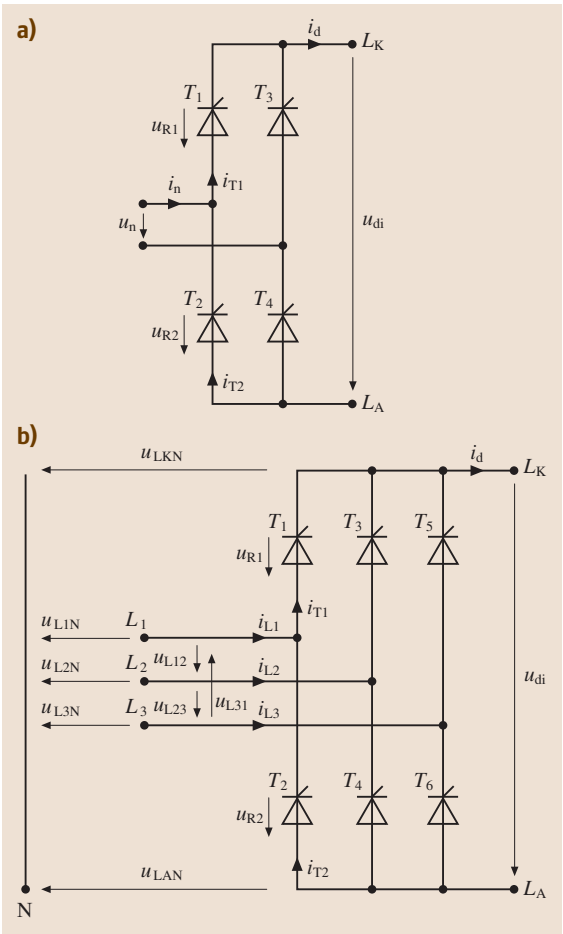


Fig. 17.68a,b Circuit diagrams of (a) single- and (b) three-phase controlled bridges

switches are diodes and half are thyristors. The operational principle of these circuits can be derived with the methods applied in this section. As a result, they constitute controllable rectifiers, being especially useful to limit inrush currents at power-on when charging a large capacitor in the DC link, as explained with reference to Fig. 17.66b.

AC Controllers

The direct conversion from one AC system to another, mentioned in Sect. 17.1, can be effected by combining a rectifier and an inverter as described in Sect. 17.4.2. In some cases a significantly simpler circuit is sufficient, i. e., an AC controller as depicted as a single-phase version in Fig. 17.69a. If the thyristor T_1 is conductive, a mains and load current $i_\mathrm{n} > 0$ can flow; if T_2 is con-

ductive, the same applies for $i_n < 0$. If a current $i_n \neq 0$ is flowing, the output voltage u_L is equal to the mains voltage u_n. The corresponding waveforms are depicted in Fig. 17.69b for an input voltage according to (17.169), a firing angle of $\alpha = \pi/4$, and a resistive load of R_L; without the AC controller, the mains current would be proportional to the voltage $i_n = u_n/R_L$. Using a firing angle of $0 < \alpha < \pi$, no current will flow before triggering the corresponding thyristor, while the current will follow the part of the sine wave after triggering. With a resistive load, i_n becomes zero again at the next zero crossing of u_n; the thyristor then blocks the reverse voltage, where possible trigger pulses have no effect. The intervals with $i_n = 0$ have two consequences. First, the AC controller reduces the RMS current in the load, which is its purpose. Second, it consumes reactive power from mains, in this case because of the phase angle between the fundamental oscillation of the mains current i_n and the voltage u_n, and harmonics due to the current shape which in total is nonsinusoidal; this again is undesirable. Note that the control behavior needs to be considered in a corresponding way, but separately for capacitive or inductive loads, causing a phase angle between the load current i_n and voltage u_L. The described behavior in principle also applies for three-phase AC controllers.

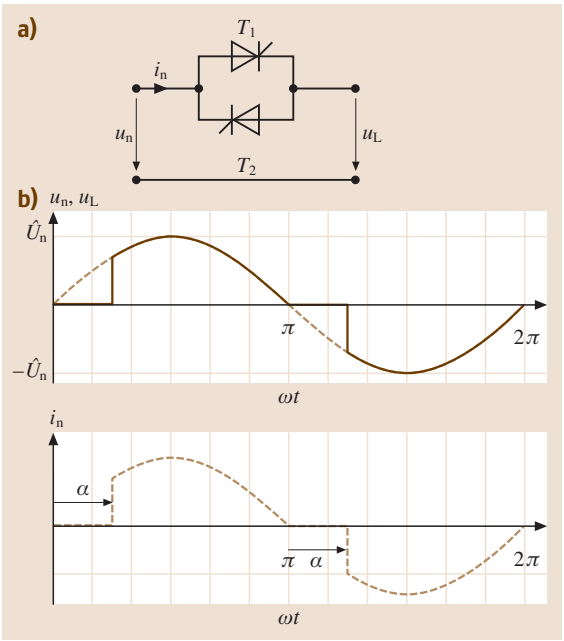


Fig. 17.69 **(a)** Circuit of a single-phase AC controller; **(b)** idealized waveforms at firing angle $\alpha = \pi/4$ with resistive load: from *top* to *bottom* – line voltage u_n (*dotted*) and output voltage u_L (*solid*) – line current i_n with label of firing angle α

Light dimmers are a typical application of AC controllers: the higher the firing angle $0 \leq \alpha < \pi$, the darker a bulb will be. The output frequency is equal to the input frequency and the nonsinusoidal shape of the output voltage u_L is no constraint in this application. Heaters can be controlled in this way as well: sometimes multicycle control is applied, where the controller is triggered with $\alpha = 0$ for several mains cycles T_n, followed by an interval without triggering. The average temperature remains smooth due to the thermal time constant of the device; no phase angle between the mains current i_n and voltage u_n occurs, however subharmonics are present in the current. In the low-power range, the pair of thyristors displayed in Fig. 17.69a may economically be replaced by one triode alternating current switch (TRIAC) with the same functionality but only one gate terminal.

17.4.4 Design Considerations

For the consideration of circuit and control theory in Sects. 17.4.2 and 17.4.3 ideal components have been assumed as stated in Sect. 17.4.1. More details need to be taken into account regarding circuit design. For this purpose, power semiconductor devices as explained with a strong physical background in [17.16] are characterized according to standards such as [17.11, 17–23]. Power losses belong to the important parameters:

- A voltage drop across the device when it is conducting leads to the following conduction losses

$$p_{VD} = u_S i_S = \begin{cases} u_{CEsat}\ i_C & \text{for IGBTs} \\ u_{DS}\ i_d = R_{DSon}\ i_D^2 & \text{for MOSFETs} \\ u_F\ i_F = -u_R\ i_F & \text{for diodes} \\ u_T\ i_T = -u_R\ i_T & \text{for thyristors} \end{cases} . \quad (17.177)$$

The conduction behavior of all bipolar devices, i. e., IGBTs, diodes, and thyristors, corresponds to diode characteristic – cf. Fig. 17.70a – while the MOSFET conducts with an almost linear on-state resistance R_{DSon} when turned on. Note that many semiconductor properties typically have a temperature dependence; the operating temperature therefore needs to be considered for the balance of losses.
An IGBT is typically in the on-state at $u_{GE} = 15\,V$,

a MOSFET at $5\,\mathrm{V} \le u_{GS} \le 10\,\mathrm{V}$, and a thyristor after a specified trigger pulse $i_G \ge I_{GT}$.

- Blocking losses

$$p_{VS} = u_S i_S \tag{17.178}$$

are often negligible. It is however essential to choose devices with rated blocking voltages higher than the applied voltage, i. e.,

$$u_Z < \begin{cases} U_{CES} & \text{for IGBTs} \\ U_{DSS} & \text{for MOSFETs} \\ U_{RRM} & \text{for diodes} \\ U_{RRM}, U_{DRM} & \text{for thyristors} \end{cases} . \tag{17.179}$$

A sufficient safety margin needs to be provided to avoid destruction by overvoltages, which may be caused by parasitic inductances in self-commutated circuits – which need to be minimized – or by voltage peaks in the mains.

An IGBT is typically in the off-state at $-15\,\mathrm{V} \le u_{GE} \le 0\,\mathrm{V}$, a MOSFET at $u_{GS} = 0\,\mathrm{V}$, and a thyristor after current has fallen below a specified limit $i_T < I_H$, unless a forward blocking voltage $u_R < 0$ is applied too early or too quickly.

- Switching losses play an important role in self-commutated circuits. Their origin can be recognized in Fig. 17.70b, where the instant t_2 when turning off the transistor T_1 of the buck chopper – cf. Fig. 17.60a – has been magnified: as stated, turn-off is effected by lowering the gate–emitter voltage. The IGBT will consequently start to desaturate and take over an increasing voltage $u_{CE} \to u_Z$ after t_{21}. Because $u_{CE} + u_R = u_Z$, the diode blocking voltage decays correspondingly. When u_{CE} reaches the DC link voltage u_Z at t_{22}, the diode voltage is consequently $u_R = 0$ and the diode can begin to conduct $i_F > 0$. When the IGBT pinches off further, the diode will thus take over the current $i_{L1}(t_2)$, which has been assumed to be virtually constant in the considered short commutation interval because of the inductance L_{L1} in the current path. At t_{23} the commutation ends with the full load current flowing through the diode $i_F = i_{L1}$. It is obvious that current is flowing through the transistor while voltage is also applied, which leads to power dissipation $u_{CE} i_C$. In the time interval $t_{21} \le t < t_{23}$ an amount of energy $E_{off} = \int_{t_{21}}^{t_{23}} u_{CE} i_C \mathrm{d}t$ is dissipated in the transistor. While the power levels may be considerable, the time intervals are typically rather short, in the range of tens of nanoseconds to few microseconds.

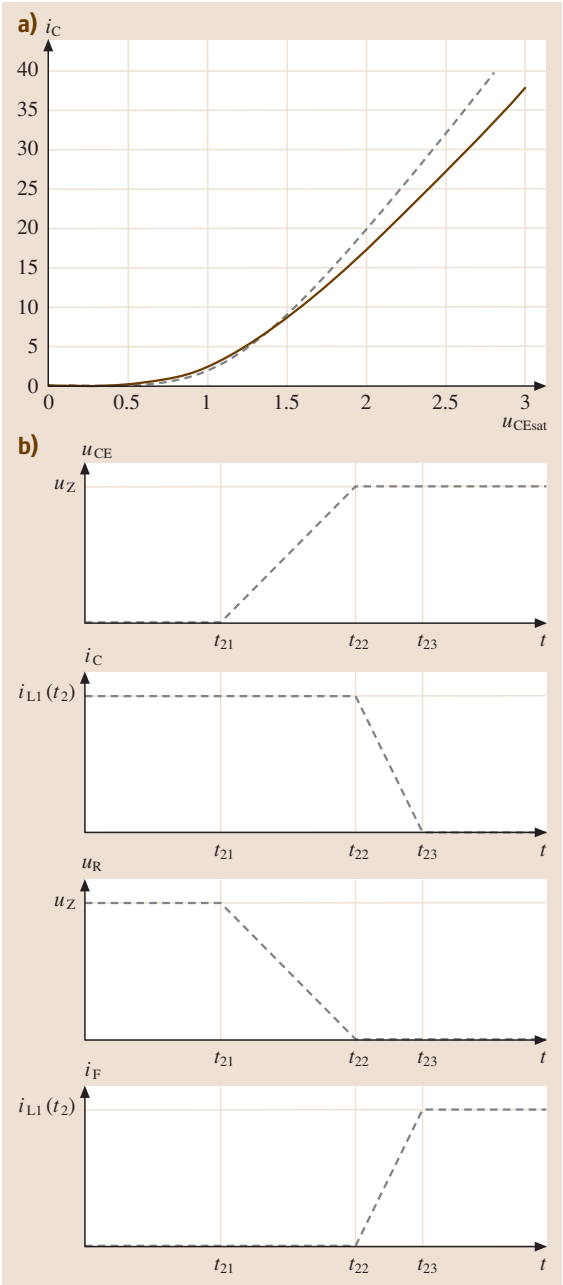


Fig. 17.70 **(a)** Exemplary conduction characteristic of an IGBT at gate–emitter voltage $u_{GE=15\,V}$ and junction temperatures $T_J = 25\,°C$ (*dotted*) and $T_J = 125\,°C$ (*solid*); **(b)** simplified waveforms of commutation in a buck chopper: from top to bottom – collector–emitter voltage u_{CE} and collector current i_C of IGBT – reverse voltage u_R and forward current i_F of the diode

The turn-on energy E_{on} of the transistor can be determined in a corresponding way. It is strongly influenced by the related diode turn-off, which is delayed by the reverse recovery effect. This is the reason why there is a variety of power semiconductor diodes with different switching characteristics; fast-recovery diodes are required in self-commutated choppers.

The fact that device switching times are not infinitesimal will in practice require the introduction of a delay between the turn-off of one and the turn-on of the other transistor in a phase leg (Fig. 17.61) thereby securely avoiding the short circuit of the bridge.

Note that besides the considered case – called hard switching – soft switching may apply, where the voltage across or the current through the device is kept at zero during switching operation; this can be achieved with resonant loads or suitable snubber networks such as in [17.24].

Assuming that the thermal time constant is much longer than the period T_P, losses can be averaged over one switching period

$$P_V = \frac{1}{T_P}\left(\int_{T_P} (p_{VD} + p_{VS})\,dt + E_{on} + E_{off}\right) . \tag{17.180}$$

Power losses, i. e., heat, need to be dissipated to ambient. For this purpose power semiconductor devices are often assembled on a heat sink, which in most cases is accessible and grounded, therefore requiring electrical isolation from the power circuit. Power semiconductor modules will include isolation while conventional discrete devices need to be isolated externally, e.g., with an insulator pad. The thermal path consequently comprises thermal resistances; it is convenient to summarize all of them between the junction (where the heat is generated) and the case as R_{thJC}, between the case and the heat sink as R_{thCS}, and between the heat sink and the ambient (where the heat will be dissipated) as R_{thSA}. Heat flow can then be modeled with an equivalent circuit according to Fig. 17.71 which permits one to derive

$$T_J = (R_{thJC} + R_{thCS} + R_{thSA})\,P_V + T_A , \tag{17.181}$$

$$T_J = (R_{thJC} + R_{thCS})\,P_V + T_S , \tag{17.182}$$

$$T_J = R_{thJC}\,P_V + T_C , \tag{17.183}$$

optionally also using the alternative indication of thermal resistance from the junction to the heat sink

$$R_{thJS} = R_{thJC} + R_{thCS} . \tag{17.184}$$

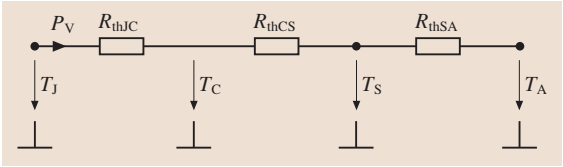


Fig. 17.71 Thermal equivalent circuit of the heat path with power P_V being transferred from the junction at temperature T_J via the case at T_C and the heat sink at T_S to ambient at T_A through thermal resistances R_{th}

With the power losses according to (17.180), (17.181–17.183) permit the evaluation of the junction temperature T_J depending on the boundary conditions T_A, T_S, or T_C; this may not exceed the rated value

$$T_J \le T_{Jmax} \tag{17.185}$$

and in some cases it should be further restricted for reliability reasons. In this way it is possible to check whether a proposed design of a power section is suitable for a particular application demanding certain current and voltage ratings of the converter.

Generally speaking, the following solutions turn out to be expedient:

- Switching losses of unipolar power MOSFETs are typically lower than those of comparable bipolar IGBTs. However, the on-state resistance R_{DSon} increases exponentially with the rated blocking voltage, saturation voltage u_{CEsat} only about linearly. This makes power MOSFETs preferable for applications with a low DC link voltage up to some $u_Z \le 100 \ldots 200$ V, e.g., in vehicle electrical distribution systems, or when high switching frequencies f_T of several tens to more than a hundred kilohertz are required, e.g., to minimize transformer size in a switched-mode power supply. Care needs to be taken that the reverse diode of the MOSFET has a reasonable switching behavior if it is used. In many other applications – in particular variable-speed drives supplied by rectified mains with a DC link voltage $u_Z \ge 300$ V and moderate switching frequency $f_T \le 20$ kHz – IGBTs in combination with appropriate fast-recovery diodes have become established.
- Complementary rectifier diodes or phase-control thyristors will be used for rectification at mains frequency. While switching losses are secondary, the insertion of inductances between the mains and a B2 or B6 bridge is recommendable at higher power ratings to limit the rate of change of cur-

rent during commutation and thus the related mains disturbance. Device surge current ratings may be important for power-on and the occurrence of mains voltage peaks.

The aforementioned control algorithms can be implemented in various ways. Some of them, e.g., for light dimmers according to Sect. 17.4.3, have been realized in integrated circuits; more complex algorithms are frequently programmed in microcontrollers or even digital signal processors. Interfaces between the power section and the control unit need to be defined: typically measurement and status signals are transferred to the control unit whereas logic levels indicating the switching states of the transistors are transmitted to the power section. The interfaces often include potential separation: while thyristors can be triggered with small transformers, the control of IGBTs or MOSFETs is mainly realized with integrated or discrete driver circuits to charge and discharge their gates; a separate power supply may be required for this purpose. Note that the potential difference between the emitters of T_1 and T_2 in the phase leg of Fig. 17.59c will vary between virtually zero (when T_2 or D_2 are conducting) and u_Z (when T_1 or D_1 are on). Because of this kind of fast transients, electromagnetic compatibility including conducted and radiated emissions should already be considered in the design phase.

17.5 Electric Drives

17.5.1 General Information

Electric drives deliver the required form of mechanical energy needed to conduct certain technical processes [17.25,26]. These drives are applied in a wide variety of industrial processes and are mainly used in:

- Machine tools
- Lifts and crane devices
- Pumps and ventilators
- Roller mills and calanders
- Valves and gate valves
- Positioning facilities and robots

Additionally, they are used in vehicles, particularly railed vehicles.

Electrical drives have the following functions:

- Control of torque (forces) and angular speed (speeds) to be compatible with machine operation and technological processes
- Optimization of machine operation to achieve technical criteria
- Insurance of minimal power losses during energy conversion

Rotating electrical machines such as induction, synchronous, direct-current machines, and their improved forms are used in these drives. For some purposes, a linear motor can also be used.

The drive device structures are addressed to technological processes:

- The electrical machine can be connected directly to the power supply for operation at fixed frequency and voltage.
- If control and regulation are required, controllable energy supply units must be used. These units mainly consist of power electronics.
- For speed, position, or torque control with closed-loop schemes the principles of cybernetics are applied.

It is therefore obvious that the field of electric drives is multidisciplinary (Fig. 17.72).

Steady-State Operation

In steady-state operation, the drive device produces a constant torque at a constant angular speed [17.26]. The intersection point between the drive characteristic and load characteristic must ensure stability for the whole drive system.

The various characteristics for these machines are often approximated by an ideal constant or quadratic, and rarely linear, relation between the torque produced M_L and the angular speed $\Omega = 2\pi n/60$ (with n in rpm). In the starting range of the machine there are usually particular variations from the ideal characteristics which mainly result from the breakaway torque. Examples of the torque characteristic for working machines and drive devices are presented in Fig. 17.73.

At constant voltage drive devices can be described by three typical characteristics in relation to $M(\Omega)$ (Fig. 17.73b):

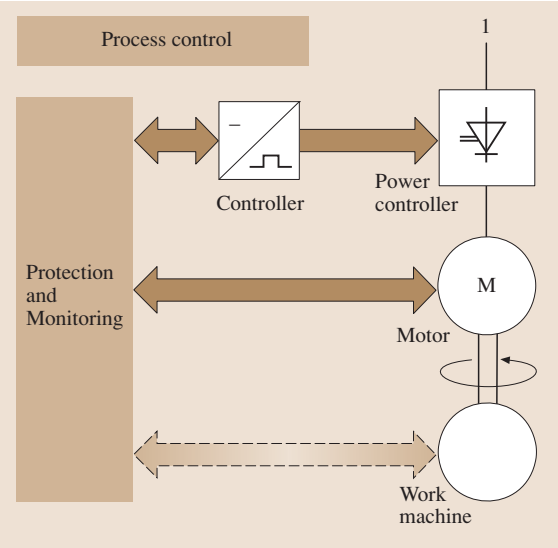


Fig. 17.72 Control schema for drive devices: 1 – power supply [17.1]

- The synchronous characteristic (for synchronous machines)
- The shunt-wound characteristic (for direct-current motors with constant flux or asynchronous machines)
- The series characteristic (for series commutator motors for DC and AC currents)

In some drives, alternating moments can additionally appear. For example, a single-phase motor can produce alternating torque at twice the fundamental frequency. When the drives are supplied through a thyristor converter, the alternating torque has a frequency based on the pulse number p of the converter. When designing a drive device, resonance phenomena produced by these types of torques must also be taken into account.

Starting Period

Usually asynchronous machines with a short-circuit rotor are switched on directly to the power supply. The short-circuit current of large three-phase machines can be reduced by using the star or delta start-up method. In the star circuit the voltage is $1/\sqrt{3}$ times the voltage in the delta circuit. Therefore, the current flowing through the branches is also $1/\sqrt{3}$ times the current flowing in delta circuit. As a result the values of the consumed power, the line currents, and the torque of the machine are reduced by 1/3 compared to the delta connection. In

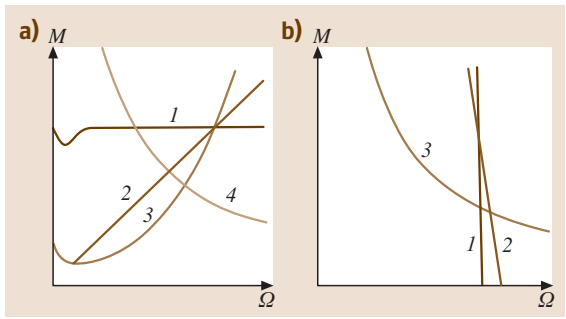


Fig. 17.73a,b Angular speed – torque characteristics of **(a)** working machines, 1 – M_L = constant; 2 – $M_L \sim \Omega$, 3 – $M_L \sim \Omega^2$, 4 – $M_L \sim 1/\Omega$, PL = const and **(b)** drive devices 1– synchronous characteristic, 2 – shunt-wound characteristic, and 3 – series-wound characteristic [17.1]

some cases, the machine can be started with the help of a starting transformer, which allows the voltage level to be increased until the machine has run up completely (Fig. 17.74).

For low-power electric machines the so-called soft-start techniques are used. These methods are applied in order to protect shafts and gears from torque surges. For three-phase rotating machines a converter that generates a linearly variable magnitude and frequency of the voltage at the clamps of the drive is used. A soft start can also be realized by the parallel connection of several resistance levels with the rotor of the machine. Changes in the resistance levels cause the machine to start-up stepwise. Instead of resistances, electronic valves can also be used.

Generally, synchronous machines are designed for asynchronous starting. Therefore, they need a start-up

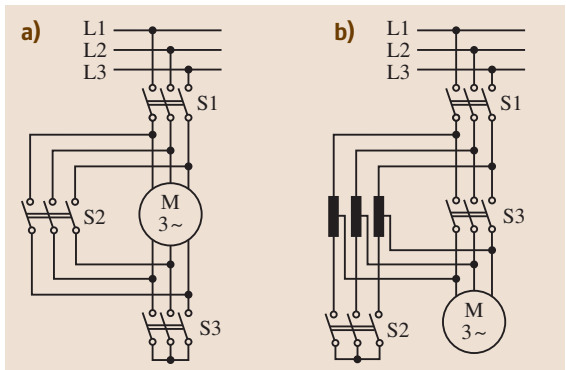


Fig. 17.74a,b Starting examples for induction machines: **(a)** star delta starting and **(b)** starting with the help of a start-up transformer

cage in the form of a damper cage. The exciting winding is shorted out with a resistance. When the rotor reaches the synchronization speed, the cage is switched off and the exciting system is switched on. Synchronization can be achieved by the use of a converter with linear frequency and magnitude control.

For DC machines, the classic starting method is the stepwise changing of a resistance at constant voltage. To reduce resistance losses during start-up, the current-leading method is mainly used. In this case, the DC machine is run up with the help of a controllable converter, which ensures maximal performance by running permanently at the current limit. This start-up system is characterized by high dynamic.

Speed Control

In controlled and regulated drive systems the angular speed is changeable with the help of a controller. There are various ways in which such systems can operate:

1. DC machines. The angular speed is controlled by suitable armature voltage regulation. By appropriate settings of the actuator, operation in all four quadrants of the coordinate system $M(\Omega)$ is possible. At speeds above the nominal speed, field control (flux control) is also used, although in this case the controlled speed range is limited. Speed control via resistance changes is also possible but it also has a lossy character. Figure 17.75 presents all of the regulation methods for DC drive devices.
2. Induction machines. The simplest method to regulate the speed of an induction machine is by changing the voltage at the clamps of the machine. This is realized with the help of a three-phase controller (Sect. 17.4.2). In this case, the idle speed is not adjustable. Because the relation between the voltage and the produced torque ($M \sim V^2$) and due to the high losses resulting from higher harmonics, this method is not suitable for high powers and wide control ranges. Thus it is only applied to ventilators with low power.
   The frequency control method is a low-loss method. In this case, the idle angular speed is adjustable. For this method the voltage needs to be adjusted as proportionally to the frequency $V \sim f(V/f)$. The breakdown torque has a constant value in the range up to the nominal voltage. If the resistance of the stator windings are taken into account, the voltage must be additionally increased for lower frequencies. If the controller is set to maximum gain, the voltage cannot be increased anymore. The frequency, typically the nominal frequency f_N of the machine, at this point is called the cut-off frequency. For further regulation the frequency is increased above the nominal frequency $f_1 > f_N$, and the machine works in the field-weakening range, which leads to a decrease of the breakdown torque. These characteristics are shown in Fig. 17.76.
   There is another conventional method for regulating the speed. By changing the number of poles the speed can be adjusted in fixed steps. The pole number of a Dahlander motor, for example, can be changed in a ratio of 1 : 2. In this case, the three-phase winding consists of six winding parts, which can work at variable pole numbers.
3. Synchronous machines. For synchronous machines, the angular speed control can only be realized through variation of the frequency and simultaneous adjustment of the voltage. The power performance of the drive device is restricted by current and voltage limitations. The speed control in the range of $0 \leq \Omega \leq \Omega_{max}$ is broken down into three general areas, see the corresponding numbers in Fig. 17.77:
   a) *Constant value of the current and flux*: linear increasing the voltage $V \sim \Omega$; from this relation the power also increases linearly $P \sim \Omega$.
   b) *Field-weakening range*: at constant voltage and current; the power level is maintained as the torque decreases.
   c) *Minimal flux* Φ_{min}: both the current and power are decreased.

Torsional Vibrations

Torsional vibrations can arise due to variable moments, rapid load changes, and short circuits. Due to possible resonances in the electromechanical system, a lot of undesirable effects can appear. In order to investigate the dynamic requirements for shafts, couplings, and gears, computational methods are used in the design stage. For this purpose, the mechanical part of the system is simulated as a multiple-mass system.

If necessary the effects of the mechanical alternating torque on the electromagnetic moment need to be taken into account.

Electric Braking

In electric drive devices, electrical as well as mechanical braking is possible. To achieve this, a reverse torque is produced by operation in the second quadrant of

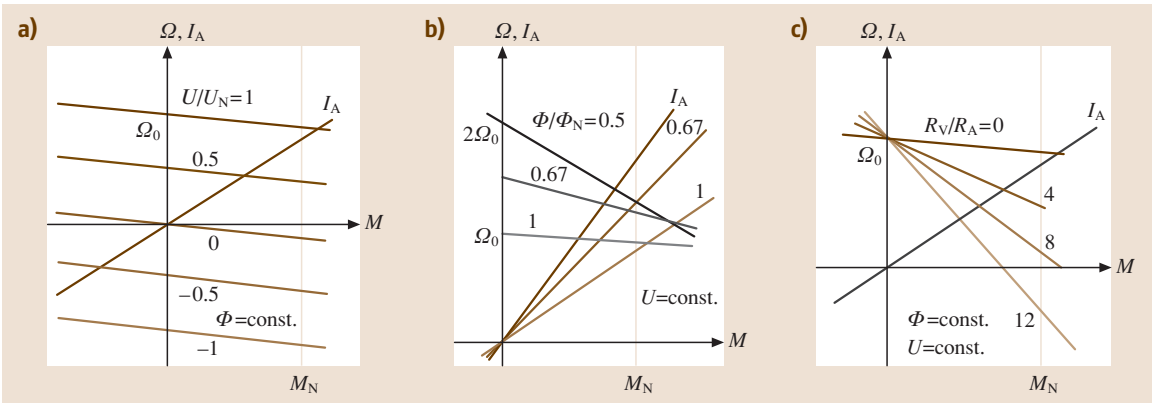


Fig. 17.75a–c Control characteristics for DC machines: **(a)** armature voltage regulation; **(b)** flux control; **(c)** resistance control

the $\Omega(M)$ characteristic. There are various methods to brake a machine:

- Regenerative brake. The braking energy flows into the network. If suitable controllers are available this is a beneficial method, which can be applied to DC (for example with self-excitation) and asynchronous machines.
- Resistance braking. During a separation from the supply system, the drive can be braked by switching certain levels of resistances to the drive. This is a classical braking method and is used mainly for DC machines. However, this method does not brake the drive all the way down to zero speed. In order to do that, either mechanical or magnetic braking is needed.
- Reverse current braking. In this case, the supply voltage is redirected, and as a result the motor tries to change the direction of rotation. If the machine is not stopped by switching off the supply voltage, the working point of the drive device will go into the third quadrant of the $\Omega(M)$ characteristic. This type of braking is used for DC current and asynchronous machines with the help of additional resistance connected to the machine. It is characterized by high energy losses because not only the braking energy but also the energy taken from the grid is converted into heat at the resistances.
- Constant-current braking. This is a special braking method used for three-phase machines. The machine in this case acts as a generator feeding a resistive load. To realize constant-current braking, the windings of the AC motor are excited with a constant current.

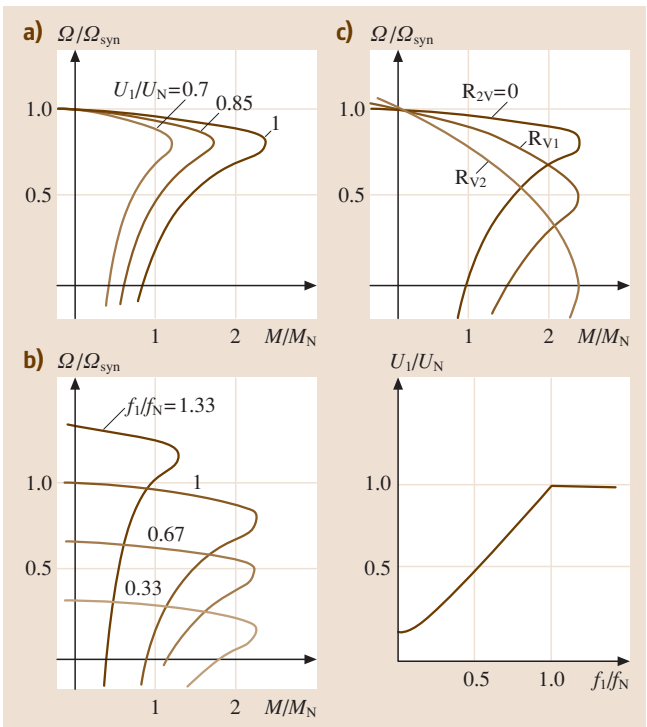


Fig. 17.76a–c Control characteristics for an asynchronous machines: **(a)** voltage control; **(b)** frequency control **(c)**; resistance control in rotor circuits

17.5.2 Direct-Current Machine Drives

Drives with Line Commutated Converters

In Fig. 17.78a a common structure for a DC-current converter is illustrated. This structure makes it possi-

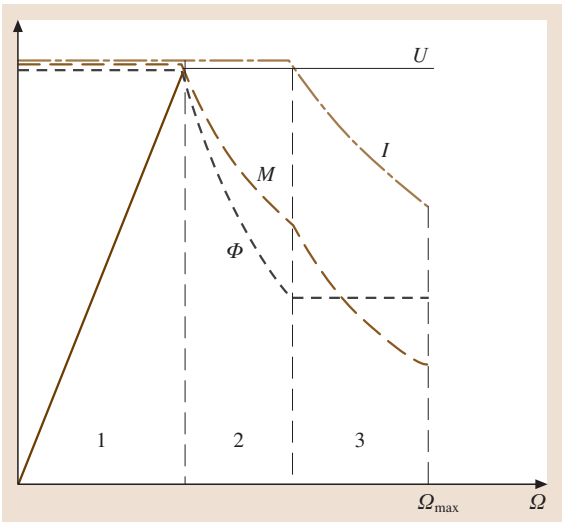


Fig. 17.77 Operation area for drive devices

ble to operate the device in two quadrants (the first and fourth). In the first quadrant the motor serves to drive the machine, and in the fourth quadrant the motor brakes the machine. An advantage in terms of reactive power compensation can be achieved by two series-connected converters (Fig. 17.78b). The equivalent circuit (Fig. 17.78c) is only valid for the continuous current mode. In this circuit, the supply system consists of the DC component, containing higher harmonics. The drive is interpreted as a DC voltage source (V_d) connected with linear inductance and resistance (L_{AM} and R_{AM}, respectively).

If a machine works in the impulse current mode this equivalent circuit is no longer valid. In this case the inductance and resistance have to be replaced by a fictitious much higher resistance. During the transition to the impulse current mode the system changes its structure. Therefore only adaptive methods can help to achieve suitable working conditions for a controlled drive device.

In steady-state operation at constant flux, the $\Omega(M)$ characteristics are similar to the load characteristics of rectifiers and inverters.

By using reverse drives, the device can be operated in all four quadrants. In order to change the drive direction, a reverse current in the armature circuit or reverse flux direction is needed. Due to the relatively long time constant of the magnetic flux, control based on changes in the current direction is preferred. To realize reverse operation of the machine a reverse converter is employed.

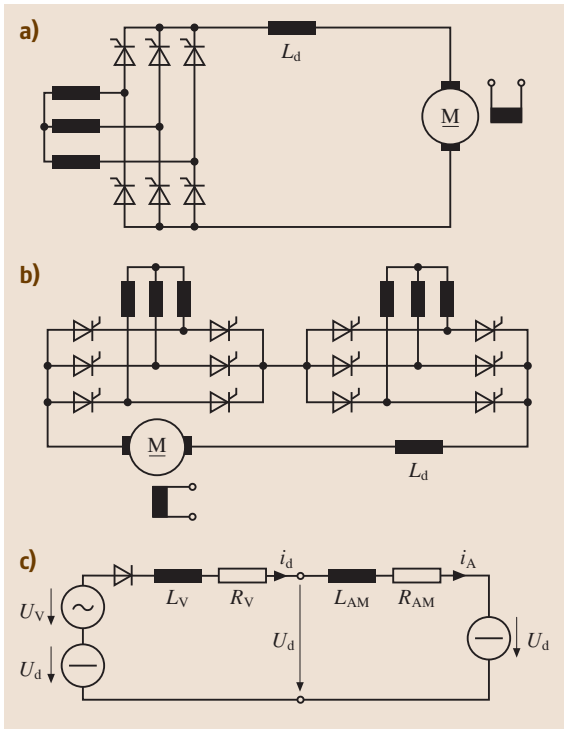


Fig. 17.78a–c DC drive device for two-quadrant operation: **(a)** circuit with current converter; **(b)** circuit with two current converters; **(c)** equivalent circuit

The commonly used structure for four-quadrant operations is two antiparallel-connected traditional controlled converters. There are two main versions to achieve reverse operation: cross-connection circuits and circuits with no DC-link current. When one converter operates as a rectifier, the second must work as an inverter. In this case the DC voltage must be higher than or equal to the voltage supplied by the second converter.

Drive Control

In drive engineering the application of control theory is indispensable [17.27]. There are two main operating principles for control systems: continuous control, which is based on analogue signals, and discrete control, which is related to digital control systems. Because of the recent progress in the field of microprocessor technologies it can be assumed that all new control systems will be based on digital processing.

To design the control system of the drive device, the following steps must be taken.

- *Modeling and system analysis*. The drive device can be described with the help of differential equations or equivalent block structures. To simplify the description and solution methods, the Laplace transformation is often used for linear time-invariant systems. The Laplace transformation makes it possible to describe the system in the form of transmittance functions in which all connections between the input and output signals are reflected in algebraic form. By using the transmittance function the relation between the signals in the frequency domain in which the dynamical properties are illustrated can be presented (Bode diagram). Time-discrete systems are described in the z-domain.
- *The controller design*. For this purpose the frequency response and root locus method are executed. The study of the stability of the system plays a very important role. For controller synthesis, a suitable parameterization is performed on the basis of optimization procedures.

If the control is based on comparison of the variance a state regulator is used. In some cases the actual value of the controlled parameter is not accessible, which is why an observer-based controller is deployed.

In many applications controllers with cascade structures are used. Therefore each controlled parameter has its own control loop. The loops are brought together in a prioritized order where the higher priorities are superposed on the lower ones.

Speed Control

The control variable of a speed control is the angular speed, which has to follow the command variable [17.28]. The control path consists of the actuator (converter), the motor, and the load. As the disturbance variable, the load torque appears. The control system is characterized by a closed-loop structure, where the control deviation, the difference between the reference and actual value, affects the control path. In terms of speed control there is another quantity to regulate, namely the current.

The dynamic properties of the controlled system are evaluated based on the transient behavior after a stepwise change of the command variable. Figure 17.79a illustrates the step response of the controlled value. It shows that the actual value has damped characteristics and exceeds the tolerance range. The rise time t_r specifies the time the control variable needs to reach the tolerance area W_T for the first time. After the transient time t_t is reached, the control variable does not leave the tolerance range again. Another characteristic term is the maximal overshoot Δ_O of the controlled variable during the transient response.

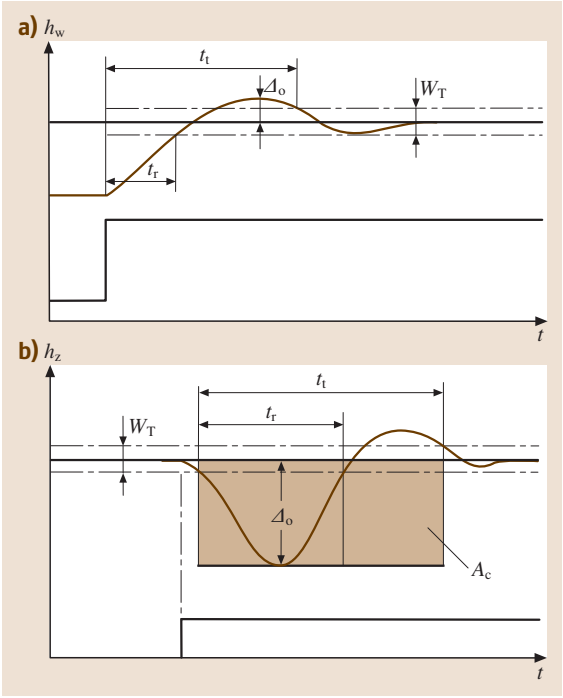


Fig. 17.79a,b Behavior of the regulated quantity by **(a)** command action; **(b)** disturbance action

A common performance index of the speed control is the quadratic form which is the integration of the quadratic control deviation. Occasionally the control area A_C, as shown in Fig. 17.79b for the disturbance variable, is used as well.

The continuous angular speed control of a DC drive is the classical representative of a cascade structure. In this case the speed control loop is subordinated by a current control loop. If necessary a position control loop can be overlaid on the speed control loop.

In Fig. 17.80, the DC machine is described by blocks 1–4. The input variables are the armature voltage v and the load torque m_L. Armature circuit 1 is simulated by a first-order time-delay element with time constant T_A. If the flux has a fixed value, the torque produced is proportional to the armature current 2 (the proportional element). The angular speed ω arises from integration 3 of the acceleration torque m_b with mechanical time

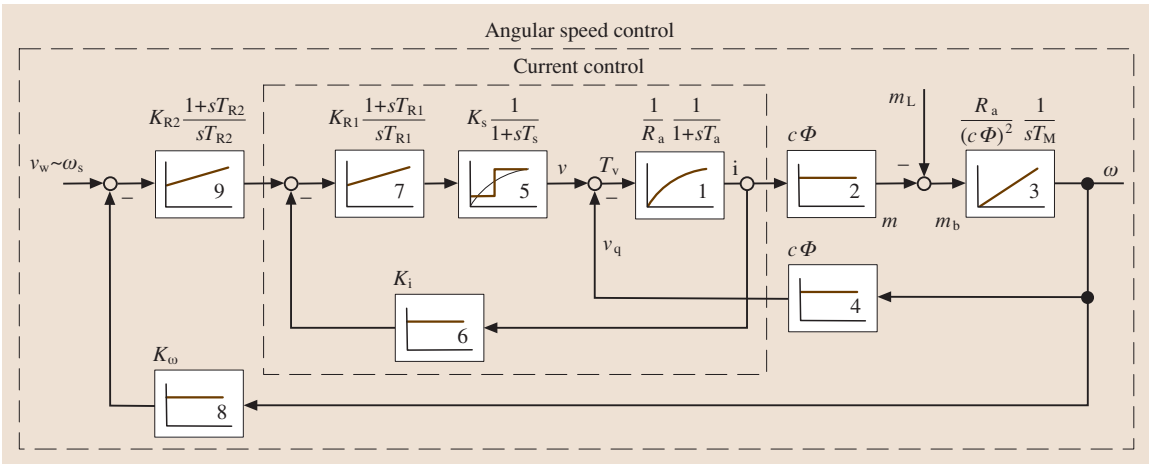


Fig. 17.80 Control loop for a DC drive device [17.1]

constant T_M. The rotating voltage v_q is created by the proportional element 4 and fed back into block 1.

The converter is reflected by an amplifying dead time element 5, which for simplification can be treated as a first-order time-delay element. Blocks 6 and 8 represent measurement equipment for angular speed and armature current, and have proportional character. The drive exhibits an angular speed control with a subordinated current control. In such a controller PID or PI controllers are usually applied. In this example, two PI controllers (blocks 7 and 9) are included in the design. The input variable for the speed controller is the control deviation $(\omega_s - \omega)$. The output value is the command variable of the current controller. The output from the controller is often limited in order to eliminate operation overload.

The required parameters of the PI controller can be set with the help of certain rules [17.1]. These procedures are explained for the given example. First, the current loop is optimized and then the overlaid controller is adjusted. In these two stages each control loop is summarized into a control block as shown in Fig. 17.81 and expressed as its step response $F_k(s)$.

In order to achieve an optimized command action of the controller the absolute value of the frequency response is required to be approximately 1, preferably for a high frequency range. This optimization is called the amplitude optimum. In this case the current control loop is optimized for the command action. By disregarding the reaction of the angular speed on the armature circuit, the transfer function of the opened loop can be expressed as follows

$$F_{k1}(s) = K_{R1}\frac{(1+sT_{R1})}{sT_{R1}}\frac{K_s}{1+sT_s}\frac{1/R_A}{1+sT_A}K_i\,. \tag{17.186}$$

The algorithms of the amplitude optimum lead to the following rules for the controller

$$T_{R1} = T_A; \quad K_{res} = K_{R1}K_S K_i 1 R_A = \frac{1}{2}\frac{T_A}{T_s}\,. \tag{17.187}$$

It becomes clear that the time constant T_A is compensated by T_{R1}, which is why the following transfer function for the closed control loop can be assumed

$$F_{w1}(s) = \frac{1}{K_i}\frac{1}{\left(1+s2T_s+s^2 2T_s^2\right)} \approx \frac{1}{K_i}\frac{1}{(1+s2T_s)}\,. \tag{17.188}$$

The last approximated transfer function is used to obtain the controller parameter for the speed control loop.

For this purpose the following function of the open loop can be applied

$$F_{k2}(s) = K_{R2}\frac{1+sT_{R1}}{sT_{R2}}\frac{1}{K_i}\frac{1}{1+sT_\Sigma}\frac{R_A}{(c\Phi)}\frac{K_\omega}{(sT_M)}\,, \tag{17.189}$$

where $T_\Sigma = 2T_s$.

The control system shown in Fig. 17.80 is generally adjusted to five a symmetrical optimum, which simultaneously gives an optimized disturbance response and short rise times after a stepwise change in the command

Table 17.6 Parameters of second-order control loop [17.1]

Methods	Amplitude optimum	Symmetric optimum	
		Without smoothing	With smoothing
Rise time t_r/T_s	4.7	3.1	7.6
Transient time t_t/T_s	11.0	11.0	14.0
Overshooting Δ_0	0.043	0.434	0.08

variable. The symmetrical optimum leads to the rules

$$T_{R2} = 4T_\Sigma,\ K_{res} = K_{R2}\frac{K_\omega}{K_i}\frac{R_A}{c\Phi} = \frac{1}{2}\frac{T_M}{T_\Sigma}\ . \quad (17.190)$$

As a result of the above approach the following transfer function is obtained

$$F_{w2}(s) = \frac{1}{K_\omega}\frac{1+s4T_\Sigma}{\left(1+s4T_\Sigma+s^2 8T_\Sigma^2+s^3 8T_\Sigma^3\right)}\ . \quad (17.191)$$

In the given example, the DC drive is considered a time-invariant system and the control system is designed continuously.

In practice, to avoid overshooting, additional smoothing of the command variable is carried out.

From (17.186–17.189) the step responses after voltage surges (command action) and torque steps (disturbance action) are determined. The characteristic values of the step responses are the rise time, the transient time, and the transient overshoot (Fig. 17.79). These values are shown in Table 17.6 for a second-order PI controller.

17.5.3 Three-Phase Drives

Drives with Three-Phase Current Controllers

An asynchronous drive with a connected current converter is adjustable within a certain angular speed range. The effective value of the voltage at the clamps of the motor is controlled by phase-angle modulation. In order to achieve a stable working point within an acceptable speed range, it is preferred to use an asynchronous machine with a resistance rotor.

The working range is limited by the maximum allowable current. The high slip values cause high rotor losses which can barely be dissipated through the self-cooling mechanism at low motor speeds. Compared to operation with a sinusoidal voltage supply the motor

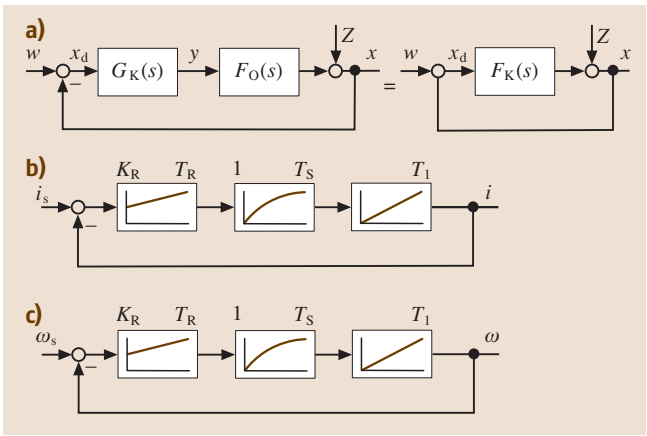


Fig. 17.81a–c Structure for controller modeling: (a) standard loop; (b) current loop; (c) angular speed loop [17.1]

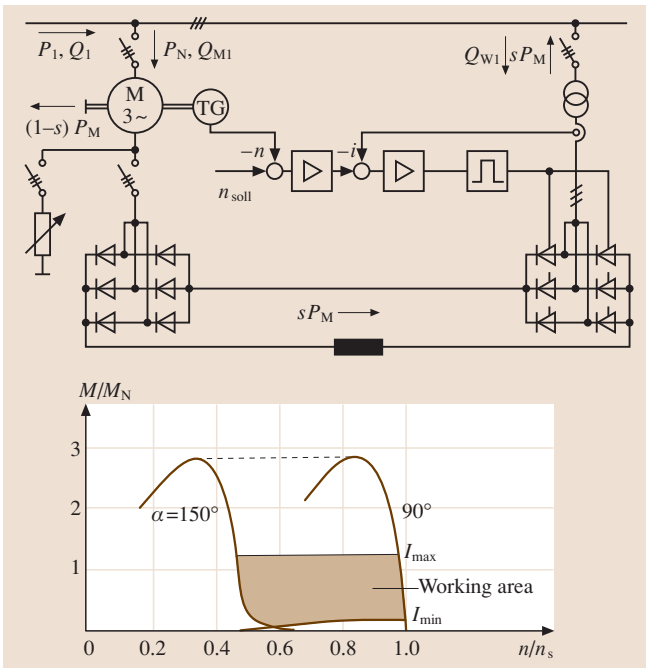


Fig. 17.82 Subsynchronous current cascade

produces additional losses due to the higher harmonics. Because of these disadvantages, this type of drive system is only used for low-power systems.

Subsynchronous Static Kraemer System

For large drive devices with a limited angular speed range of up to 2 : 1 a subsynchronous static Kraemer system can be used.

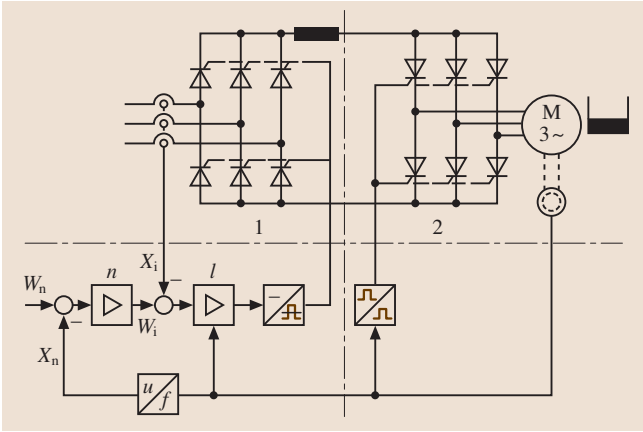


Fig. 17.83 Electrical machine with a current converter

In Fig. 17.82 the schema for a Kraemer system is shown. To realize such systems induction motors with slip-ring rotors are employed. The slip power of the rotor is fed into the power network by an indirect converter. The stator side of the motor is directly connected to the network. Disregarding the losses, the system can be analyzed in the following way:

- Real power. The electric machine consumes the power P_{M1} from the network. If the drive runs with slip s, than the mechanical power output is expressed as $P_{M2} = (1-s)P_{M1}$. The rest of the energy is given off as slip power through the indirect converter which is regulated in an angle range of $90° \leq \alpha \leq 150°$. In such a case, the power taken from the network is only $P_1 = (1-s)P_{M1}$.
- Reactive power. The machine needs reactive power Q_{M1} to operate. In addition, the inverter needs reactive power as well. The required reactive power is expressed as $Q_{W1} = V_{di} I_d \sin\alpha$ for the fundamental frequency. As a result the network has to supply the reactive power $Q_1 = Q_{M1} + Q_{W1}$.

It can be seen that the $\cos\varphi_1$ value of the drive becomes smaller as the rated power of the inverter becomes bigger; it is dependent on $P_{smax} = s_{max} P_{M1}$ and therefore also on the regulation range. Consequently, the subsynchronous system is not attractive for a wide regulation range.

Converter-Fed Motor

Common converter-fed motors are synchronous machines which are supplied by an indirect converter. The controlled rectifier ensures a suitable current level; this converter is commutated by the network voltages. The inverter forms the current waveform for the machine. Here, commutation takes place with the help of the induced machine voltages (Fig. 17.83)

Converter-fed motors are designed for power in the range of 1000 kW–20 MW.

During start-up (rotational speed up to 5–8% of the rated speed) the machine cannot ensure the commutation voltage for the inverter. In this case, additional methods are needed. The most common method is known as link-circuit pulsing. The link current is pulsed by the rectifier and provided cyclically to the machine winding by the inverter.

Control of Three-Phase Drives

In compensated DC machines, the excitation flux and armature do not influence each other. In induction machines, however, the flux produced by the stator and flux generated by the rotor are coupled. Voltage changes at the clamps of the stator cause changes in both components of the current, the flux-creating and the torque-producing component.

In DC machines, the excitation flux (d axis) is always perpendicular to the armature flux (q axis). In asynchronous machines, the flux is produced by the magnetization components of the stator current. The flux phasor exhibits an angle φ_s with reference to the static α–β coordinate system. Decoupling is achieved when the rotating system is linked to the synchronous rotating flux phasor. Now a rotating l–m coordinate system is created. This practice is called field orientation. In this system the stator current phasor exhibits the components i_m, i_l which are denoted as the flux- and moment-creating components, respectively. The torque is proportional to the product of Ψ_r and i_l. Control with the help of the field-orientating method gives good results in terms of the dynamic properties, as is known from DC machines. In order to estimate the control parameters for such drive devices, the measurable parameters are applied to partial models of the machine and transformed to the rotating coordinate system.

17.6 Electric Power Transmission and Distribution

17.6.1 General Information

Electric energy is produced in power plants and consumed by customers located at various distances from the plants. Power networks and related equipment are needed for the transmission and distribution of electric energy from the plant to the customer. Thus the following devices are used:

- Overhead lines and power cables
- Transformers
- Switching stations

These devices also include measuring transformers, protection units, and relays. Power networks are distinguished according to their purpose, with different voltage levels and structures.

The usual nominal voltages of transmission systems are high voltages of 110, 220, and 380 kV. In special cases the three-phase current is transmitted at an extra high-voltage level of up to 765 kV. Distribution systems are operated at medium voltages of either 10 or 60 kV, while low-voltage supply grids usually work at nominal voltages below 1 kV, i. e., 230 V (single phase) or 400 V (three phase).

The related operating frequency for all voltage levels is generally 50 Hz in Europe and 60 Hz in North America and Japan. Other frequencies are also used for special applications such as railway supply systems in Germany, Austria, and Switzerland, which operate at $16\frac{2}{3}$ Hz.

When choosing which basic voltage to use technical and economical factors have to be considered. In the case of long-distance energy transmission voltage stability plays an important role. Additionally, the isolation of the network components must always be able to bear the short-circuit current.

Another possibility for the transmission of electric energy is the use of the high-voltage direct-current transmission system (HVDC) at a few hundred kV. This technology is applied either for long distances or for the coupling of two systems with different frequencies and decoupling the short-circuit power of both systems to protect network components in the case of faults.

Figure 17.84 shows an example of an urban electric power network structure. The voltage levels for transmission are 380 kV and 110 kV, connected as closed loops. This voltage is then stepped down in transformer stations from 110 kV to the distribution level of 10 kV. This medium-voltage network is connected in open-loop circuits. This is advantageous in the case of faults within the medium-voltage network because the sectioning point can easily be moved within the ring and the customers are quickly reconnected to the grid.

If the number of customers connected to a given grid increases, more energy will be transmitted. Therefore, the network capacity must be increased. By increasing the network capacity and the amount of network interconnections the short-circuit power also increases. Ideally, the short-circuit power is the product of the short-circuit current and the rated voltage. This value does not appear in reality but is used for the dimensioning of network components. An increased short-circuit power can be dangerous for network components if a fault occurs because dynamic and thermal effects lead to their destruction. To limit the short-circuit power several actions have to take place. Some of these concern the neutral-point treatment, which has an influence on the resulting short-circuit current.

Networks are generally distinguished according to the neutral-point treatment of the transformers. There are isolated neutral points, impedance grounding, and direct grounding. (Figs. 17.85–17.87). When faults occur – especially single-phase faults – some neutral-point structures can continue operation, whereas some others have to be disconnected immediately. Which structures can be used and which must be disconnected depends on the type of neutral-point treatment deployed.

The fault-to-earth current is determined mainly by the earth capacitance and must not be higher than 35 A to self-extinguish. During a fault, the network can continue to operate when an isolated neutral point is used. (Fig. 17.85). The isolated neutral point can be applied to small networks with voltages below 30 kV.

By using a specific type of impedance grounding (Fig. 17.86) (e.g., an *arc suppression coil* or *neutral grounding reactor*), in the case of a phase-to-ground fault, the capacitive components failure current will be compensated by the reactor (*Petersen coil*). As a result, the residual current of the arc will be small enough to self-extinguish if it has a value below 100 A. The voltages in the faultless phases are multiplied by the factor $\sqrt{3}$ during the fault in one phase. Like the isolated neutral-point network, this type of network can also continue to operate during a fault.

Impedance grounding is used for networks with voltages below 110 kV.

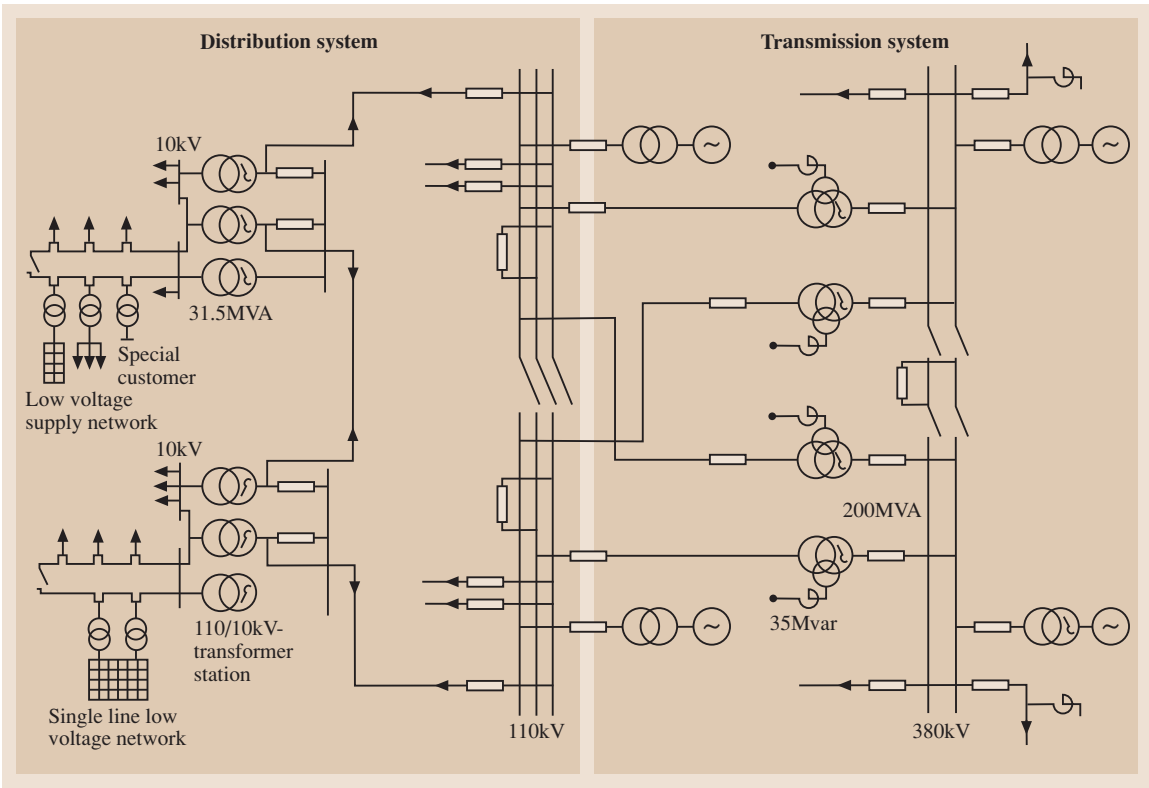


Fig. 17.84 Example of an electric power network structure

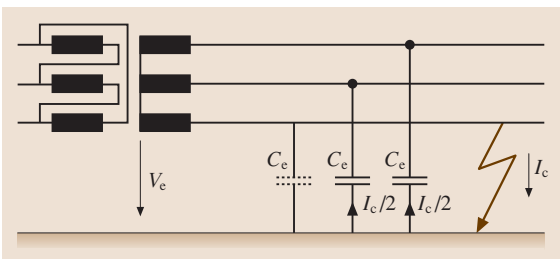


Fig. 17.85 Isolated neutral point

Direct grounding is mainly used in HV networks operating at higher than 110 kV (Fig. 17.87). Here an automatic protection system is necessary because the self-fault-clearing of the short-circuit arc cannot be ensured. Due to the small impedance of the neutral point the fault-to-earth current is a short-circuit current and therefore has to be cut off immediately. The overvoltages resulting from a fault in a direct-grounded system are a lot smaller than in the other structures mentioned above.

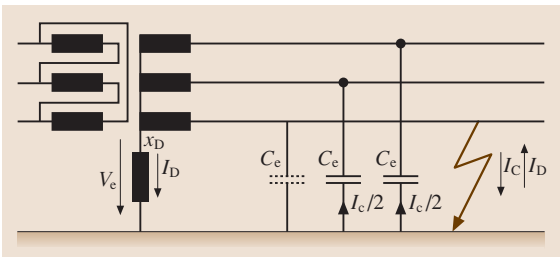


Fig. 17.86 Impedance grounding

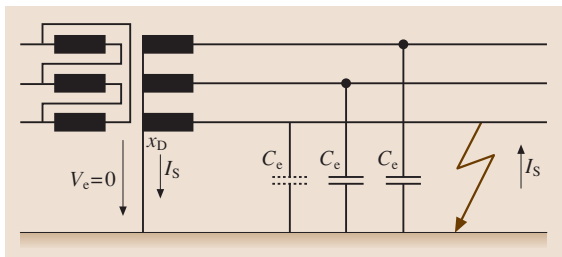


Fig. 17.87 Direct grounding

17.6.2 Cables and Lines

Electrical energy is conducted from the generator to the customer through either cables or lines. The relative losses per phase of a three-phase line can be defined as

$$\frac{\left(P_\mathrm{v}/l\right)}{S} = \frac{S}{\left(V_\mathrm{L}^2 \kappa A\right)} , \qquad (17.192)$$

where κ is the electrical conductance, A is the cross-section per phase, l is the length, S isthe apparent power and V_L is the line voltage.

This equation reveals why a high voltage is advantageous for transmitting a definite apparent power – the higher the voltage, the smaller the losses. For economical and technical reasons mainly overhead lines are used in high-voltage systems (110–380 kV) instead of cables. Cables are used up to a voltage of 110 kV (in special cases up to 380 kV) in densely populated areas, and sometimes even outside cities due to visible impact.

Line Parameters

The one-phase equivalent circuit shown in Fig. 17.88 can be used for both cables and overhead lines.

Due to their construction three-phase cables have symmetric characteristics. The three phases of an overhead line have to be twisted within a certain distance to achieve this quality.

The active resistance R depends on the length l, the cross-section A, the electrical conductance κ as well as on temperature ϑ

$$R = \frac{1}{\kappa A(1+\alpha\Delta\vartheta)} . \qquad (17.193)$$

The operating inductance L_b of a phase line can be expressed as

$$L_\mathrm{b} = \frac{\mu_0 l}{2\pi}\left(\ln\frac{d_\mathrm{m}}{r} + \frac{1}{4}\right) . \qquad (17.194)$$

where d_m is the average conductor distance and r is the conductor radius.

The operating capacitance C_b consists of the capacitance to earth C_e and the capacitance C_g between phases, which is converted from a star connection to a delta connection Fig. 17.89

$$C_\mathrm{b} = C_\mathrm{e} + 3C_\mathrm{g} = \frac{(2\pi\epsilon_0 l)/\ln d_\mathrm{m}}{r} . \qquad (17.195)$$

Cables have a remarkably higher capacitance per length C' than overhead lines. This has an effect on the charging current which is 25–40 times higher for cables, and on the earth fault current, which is more capacitive in cable networks and has to be compensated.

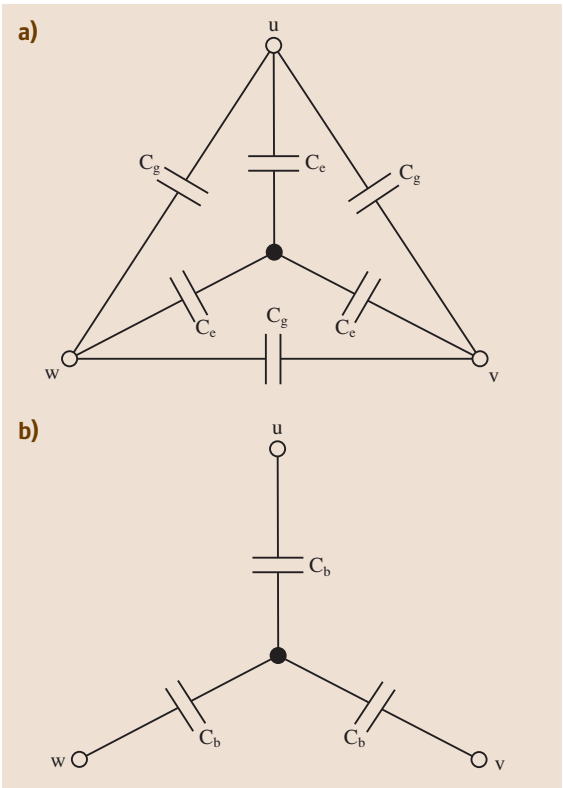


Fig. 17.88a,b Three-phase system capacitance: **(a)** part capacitance; **(b)** operation capacitance

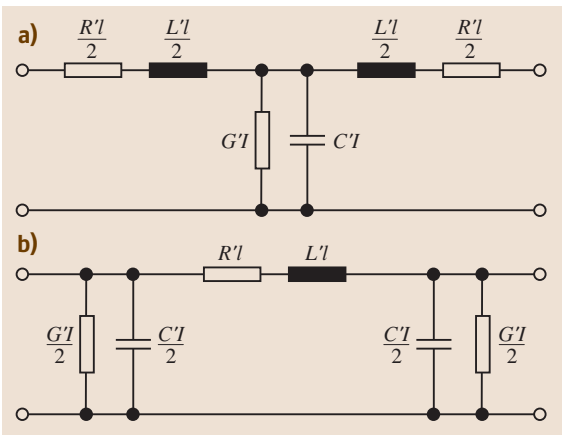


Fig. 17.89a,b Equivalent circuit of a line: **(a)** T-circuit; **(b)** Π-circuit

The specific values for resistance, inductance, and capacitance per length of a line are minimally dependent on the voltage if the operating voltage is changed

$$I_c = \frac{V_l \omega C_b}{\sqrt{3}} \,. \tag{17.196}$$

Line Simulation

Figure 17.89 shows an equivalent circuit for one line. A line can be described using:

- The resistance per unit length R'
- The inductance per unit length L'
- The capacitance per unit length C'
- The conductance per unit length G'

The equivalent circuits can have either a T or Π form.

In most cases the conductance per length G' can be neglected because $G' \ll \omega C'$.

The line operation can be described using line theory, which helps to determine the voltage and current at each position of a line. Therefore, the line impedance is needed. The impedance is generally identified as the characteristic wave impedance Z_W

$$\underline{Z}_W = \sqrt{\frac{R' + j\omega L'}{G' + j'\omega C'}} \,. \tag{17.197}$$

For a lossless line this can be reduced to a resistive value

$$\underline{Z}_W = Z_W = \sqrt{\frac{L'}{C'}} \,. \tag{17.198}$$

The current and voltage along a line can be represented as a superposition of a forward-running wave (w_f) and one reflected, backward-running wave (w_b). The voltage at the end of a line v_2 can be calculated by the summation of both waves

$$v_2 = w_f + w_b \,. \tag{17.199}$$

The current can be calculated by the difference between w_f and w_b, divided by the wave impedance

$$i = \frac{1}{Z_W} (w_f - w_b) \,. \tag{17.200}$$

The relation between the backward and forward components of the voltage is expressed as a reflection coefficient r, which depends on the terminating impedance $\underline{Z}_a$ of the line and the wave resistance $\underline{Z}_W$

$$\underline{r} = \frac{\underline{w}_f}{\underline{w}_b} = \frac{\underline{Z}_t - \underline{Z}_w}{\underline{Z}_t + \underline{Z}_w} \,. \tag{17.201}$$

The reflection coefficient for unloaded operation ($\underline{Z}_t = \infty$) is $\underline{r} = 1$ and for short-circuit operation ($\underline{Z}_t = 0$) it is $\underline{r} = -1$.

This results in a voltage of $v_2 = 2w_b$ for unloaded operation and $v_2 = 0$ for a short circuit.

The reflection coefficient is $\underline{r} = 0$ for $\underline{Z}_t = \underline{Z}_w$, in the case of complete adjustment of the terminating impedance. In this case there are no backward-running waves and as a result, the voltage at the end of the line is the same as at the beginning ($v_2 = w_f$). At this point the transmission of maximal power P_n with minimal losses is possible

$$P_n = \frac{V_L^2}{Z_W} \,. \tag{17.202}$$

This active power P_n is called the *natural power*. For 380 kV overhead lines it has a value of about 450 MW.

17.6.3 Switchgear

Switching Stations

Switchgear is a general term covering switching devices and their combination with associated control, measuring, protective, and regulating equipment. Switching stations are needed to collect and distribute electric energy.

High-voltage switching stations up to 110 kV are usually located within buildings; higher voltages require outdoor switching stations or special encapsulated switching stations based on SF_6 gas technology. Due to its electric strength it can be used in places where insulation with air would take too much space.

High-Voltage Switchgear

The three types of switchgear, each with a different purpose, are:

- A *circuit breaker* is intended to switch both load and short-circuit currents. The switching power can be calculated as a product of the rated voltage and the switching current. It is dimensioned to avoid damage due to the short-circuit current's dynamic and thermal effects on generators, transformers, switching stations, and lines. It restores quickly after a short circuit and is the most effective but also expensive form of switchgear. It

is not primarily intended for frequent operation, although newer vacuum and SF_6 circuit breakers are more suited to load-switching duties than older switchgear types.

- An *on-load switch* is intended for switching loads under normal and overloaded conditions. It is designed for frequent operation but has a limited short-circuit current-carrying and switching capability. It is therefore often backed up by fuses or a circuit breaker.
- An *air-break disconnector* provides a visible isolating distance in the open position. It has nearly no current-switching capability and is used for security of personnel, to show that a circuit is open. It is usually connected in series with a circuit breaker or an on-load switch.

In Fig. 17.90 an equivalent circuit of a circuit breaker is shown, to explain the basic principle of switching. The circuit contains the source impedance $R_s + j\omega L_s$, the switch, followed by the circuit breaker's downstream impedance $R_L + j\omega L_L$, and the fault. The shunt impedance can be neglected as it is several orders of magnitude smaller than the other series impedances involved. Initially, the circuit breaker is closed and carries the fault current $i_s(t)$. The relay protection notices a fault and initiates the opening of the circuit breaker. Figure 17.91 shows the behavior of the short-circuit current $i_s(t)$, through the circuit breaker and the voltage across the circuit breaker $v_{cb}(t)$. At time t_1, the circuit breaker contacts begin to separate and an arc ignites across the contacts. The arc is then extinguished by the particularly used arc-quenching mechanisms, which can differ between different types of circuit breakers. This involves stretching the arc and its intense rapid cooling.

The duration of the arc is rather short (≈ 10 ms) and is coupled with a low arc voltage. The characteristic waveform of the recovery voltage is also shown in Fig. 17.91. A high frequent voltage oscillation, also known as the transient recovery voltage, appears as soon as the short-circuit current passes the zero line and is cleared at time t_2. The behavior of this voltage is determined by the specific circuit parameters.

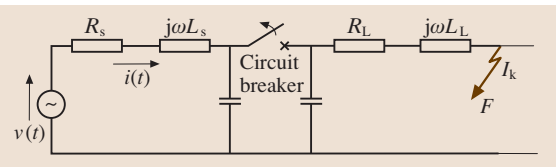


Fig. 17.90 Circuit breaker

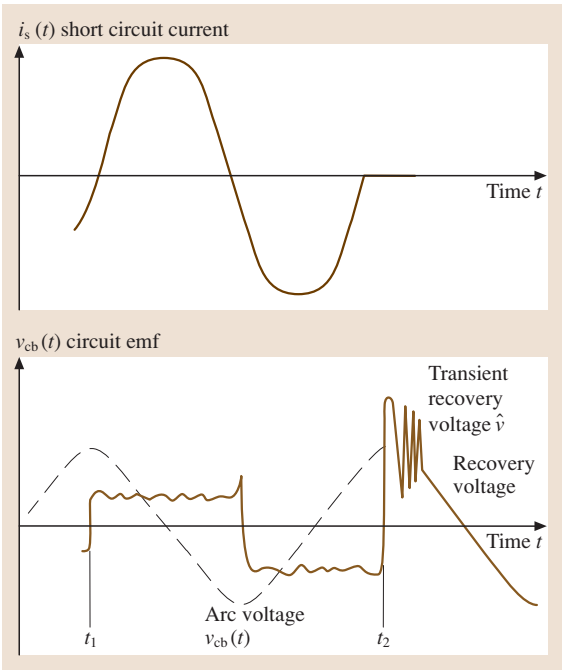


Fig. 17.91 Behavior of the short-circuit current and recovery voltage

Low-Voltage Switchgear

In low-voltage systems below 1000 V many different types of switchgear exist. They can be distinguished according to the:

1. Type of activation:
   - Manually controlled switch
   - Impulse controlled switch
   - Air-gap switch
2. Switch ability:
   - Disconnector
   - On-load switch
   - Circuit breaker
   - Motor switch
3. Intended use:
   - Control switch
   - Limit switch
   - Disconnecting switch
   - Protection switch

17.6.4 System Protection

Short-Circuit Current Protection

To rapidly clear short circuits either power circuit breakers with magnetic tripping devices or melting fuses

are mainly used. Rapid clearance is important because the short-circuit current is much higher than the nominal current and can seriously damage power system devices.

Protection Switches

In low-voltage networks the protection function is often taken over by the switch gear. When a fault occurs, e.g., an overcurrent or voltage sags, the switch gear is able to switch off the endangered part of the system. *Key switches* release the energy that was stored in a latch mechanism and use it for tripping. Other switches are moved into their starting position by a resetting force which is activated by the omission of the driving force. *Contactors* are switches whose contacts are kept in the turn-on position using a current-carrying magnet coil. Contacts within the control circuit of the contactor are responsible for tripping because their control current is interrupted. Contactors are mainly used as a motor protection switch combined with thermal overcurrent relays and instantaneous magnetic tripping, see Fig. 17.92. Tripping devices and relays can be arranged in different ways for instantaneous, delayed or time-selective cutoff of power circuit breakers.

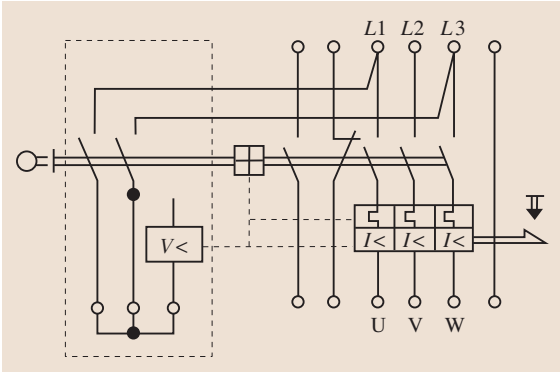


Fig. 17.92 Circuit for motor switch protection

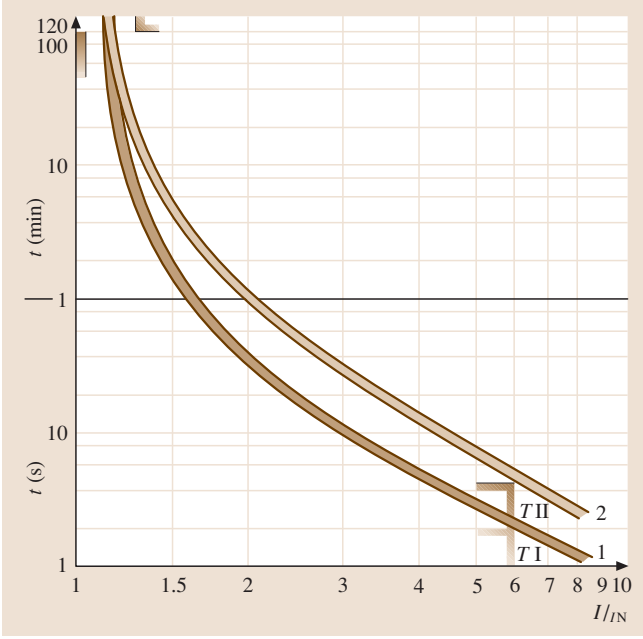


Fig. 17.93 Breakdown characteristics of overcurrent relays

Thermal Overload Protection

Conventional protection devices are bimetal tripping devices and relays. Their operation is based on the behavior of belts composed of two materials with different expansion coefficients. Bending of the bimetal is caused by internal tension due to warming as a result of current flow (Fig. 17.93).

Short-Circuit Current

When choosing devices and equipment, the dynamic and thermal demands appearing during short circuits have to be considered. Thus, and also because of possible unallowed high contact voltages, the highest possible short-circuit current has to be calculated during the design process. Also the smallest possible short-circuit current is important for the dimensioning of the protection system.

In three-phase systems several kinds of short circuits can be distinguished:

- Three-phase short circuit. This is a symmetrical short circuit, with alternating short-circuit currents depending only on the impedance of the positive-sequence system and forming a symmetrical three-phase system. This results in the highest thermal stress and dynamic operational demand for the network components.
- Nonsymmetrical short circuits. This category includes two-phase short circuits with and without contact to ground, one-phase line-to-ground short circuits, and the double-earth fault. The short-circuit currents are defined by the impedances of the positive- as well as of the negative-sequence system and possibly the impedance of the zero-sequence system.

To determine the short-circuit currents the applied voltages and short-circuit impedances have to be known.

In a simple network with one three-phase generator which is transferred into symmetrical components the subtransient voltage E'' determines the initial short-circuit current I''_s. The subtransient voltages of the generator are $\underline{E}''_1 = \underline{E}''$, $\underline{E}''_2 = 0$, and $\underline{E}''_0 = 0$, given in symmetrical components. The generator only produces a positive-sequence voltage. The impedance between the generator and the fault location is composed of impedances in symmetrical components $\underline{Z}_1$, $\underline{Z}_2$, and $\underline{Z}_0$, which are mainly inductive, so often $\underline{X}_1$, $\underline{X}_2$, and $\underline{X}_0$ are used (Fig. 17.95).

Below, the three most important cases are described:

- The maximum initial short-circuit alternating current I''_s (starting current) can be calculated using

$$\underline{I}''_{s3} = \frac{\underline{E}''}{\underline{Z}_1} \,. \tag{17.203}$$

- The initial short-circuit alternating current I''_k in a two-phase short circuit without earth contact (line-to-line fault between phase V and W) will be

$$\underline{I}''_{s2} = \frac{\sqrt{3}\underline{E}''}{(\underline{Z}_1 + \underline{Z}_2)} \tag{17.204}$$

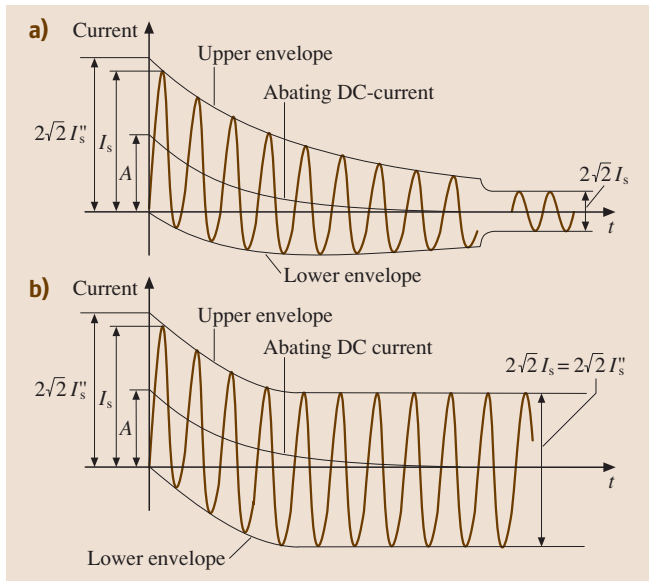


Fig. 17.94a,b Characteristic of a short circuit: **(a)** close to the generator; **(b)** far from the generator

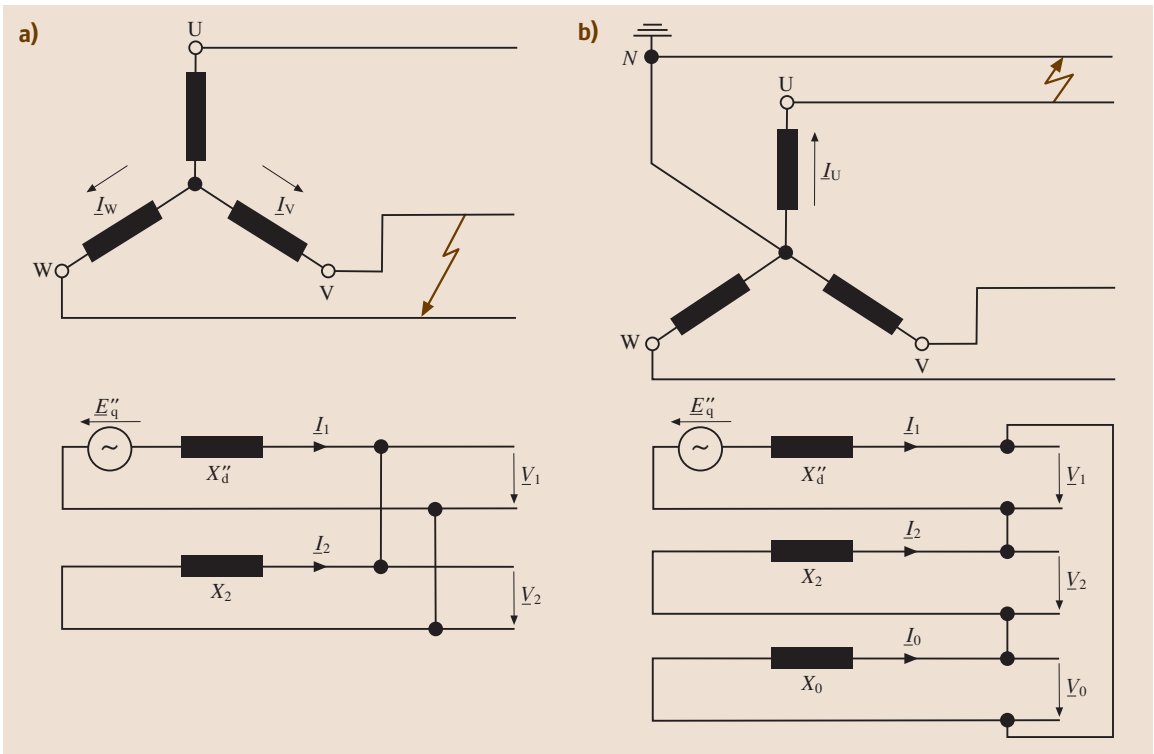


Fig. 17.95a,b Unsymmetrical generator short circuits: **(a)** two phase and **(b)** single phase

considering the fault conditions

$$\underline{I}_U = 0;\ \underline{I}_W = -\underline{I}_V;\ \underline{V}_V - \underline{V}_W = 0\,. \qquad (17.205)$$

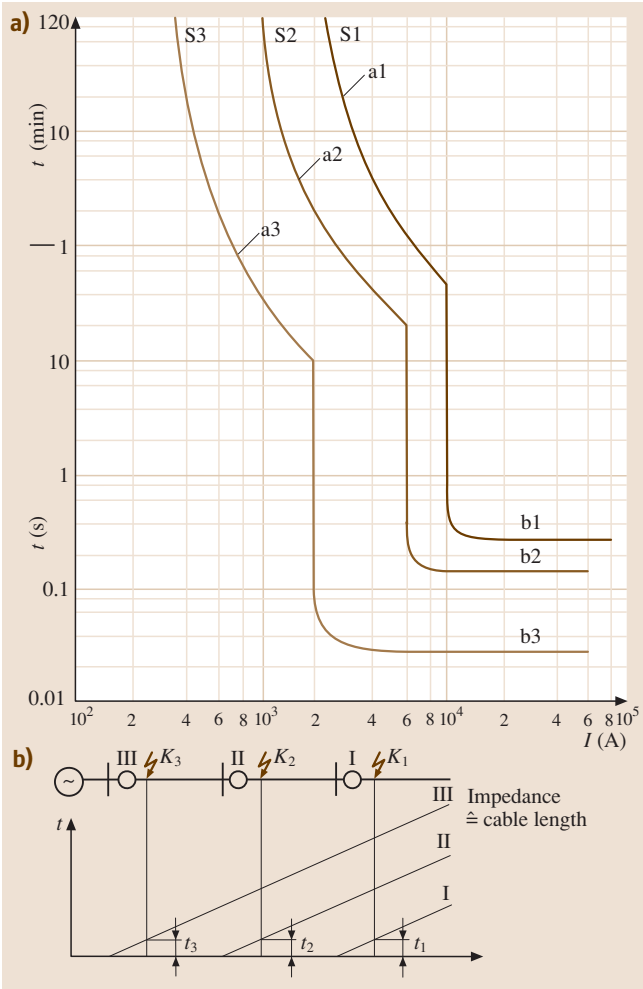
- The initial short-circuit alternating current I''_k for a phase-to-ground fault of phase U will be

$$\underline{I}''_{s1} = \frac{3\underline{E}''}{(\underline{Z}_1 + \underline{Z}_2 + \underline{Z}_0)} \qquad (17.206)$$

considering the fault conditions

$$I_V = I_W = 0 \quad \text{and} \quad V_U = 0\,. \qquad (17.207)$$

Figure 17.95 shows examples of two kinds of non-symmetrical short circuits, represented by means of symmetrical components.



The standard DIN VDE 102 defines the calculation method for any kind of short circuit using an equivalent voltage source $(cV_h)/\sqrt{3}$ implemented at the fault location, with V_h being the line-to-line voltage and $c = 1.1$ for high-voltage networks. For low-voltage networks $c = 0.95$ for the calculation of the smallest possible short-circuit current.

If the initial short-circuit alternating current I''_k is known, the peak short-circuit current can be determined using $I_S = \kappa\sqrt{2}I''_k$, with κ taking values between 1 and 2 depending on the relation R_k/X_k of the short-circuit loop.

In more complicated networks the short-circuit analysis requires the use of special software or, in former times, special analogue network models. Thus, dynamic processes can also be investigated.

Selective Network Protection

If a short circuit occurs, only the affected parts of the network should be cut off. Thus selective network protection has to be achieved, which is possible by using time grading for radial networks. This means that the time delay of a circuit breaker located closer to the fault should be smaller than the time delay of the next circuit breaker and not dependent on the current. Between several power circuit breakers connected in series, a time delay of about 60 ms is used.

Figure 17.96 shows an example of the tripping characteristics for three selective time-graded switches. The selectivity within an intermeshed network is achieved using distance protection.

At the power switch location, currents and voltages are measured and the network impedance is calculated. The tripping time increases with increasing network impedance, thus the switches which are closest to the short-circuit area will open first (Fig. 17.96b).

Alternatively, a selective high-speed switch used for automatic rapid reclosing can be used. In the case of a short circuit all switches in a current-carrying path are turned off simultaneously after 10 ms. If the short circuit only consists of an arc it is feasible for it to self-extinguish. After a rapid reclosing of 700 ms the switches will turn on again. If the short-circuit current still flows, they will open again.

For the parts of the energy system not containing generators or loads the difference protection system

Fig. 17.96a,b Selective network protection: **(a)** tripping characteristic of selective time-graded power circuit breakers; **(b)** tripping characteristics of distance protection with tripping time depending on fault location◀

can be employed. Difference protection is based on the comparison of the input and output currents, if necessary considering the related transformer ratio; failures are thereby identified and the appropriate circuit is shut off.

Oil transformers are additionally equipped with Buchholz relay protection, which responds to gas production within the transformer due to a fault.

Electric Shock Protection

Electric shock protection is necessary to prevent persons from touching voltage-carrying parts. According to IEC 61557 (DIN VDE 0100) the acceptable limit of touchable voltage is 65 V. Protection arrangements should prevent related parts of a system from carrying higher voltages under normal operation conditions as well as in the case of faults.

There are various forms of protection, including:

- *Protection by low voltage*: a maximum operating voltage of 42 V is required for devices operating at protection low voltage.
- *Protective insulation* is used to prevent too high contact voltages from bridging; this requires additional insulation.
- *Protective separation* uses an isolation transformer to disconnect the load circuit from the supplying network and therefore avoids a contact voltage between the equipment and earth in the case of body contact.

The purpose of further protection techniques is to ensure switching off the respective device in the case of isolation faults; thus an unallowable contact voltage will not appear.

These devices need to have a protective grounding conductor for:

- *Protective grounding*, in which the device's case is either connected to the earth or to grounded parts.
- *Neutral grounding* requires a connection between the device's case and a directly grounded wire, e.g., a grounded neutral wire.
- In a *protective conductor system* all respective devices of the electric system are connected to each other as well as to conducting parts of the building, the piping, and grounding electrode.
- *Voltage-operated earth-leakage protection* is able to disconnect all phase conductors within 0.2 s in the case of high contact voltages.
- The *current-operated earth-leakage circuit breaker* is able to disconnect within 0.2 s in the case of a fault current (e.g., 60 mA), thus there will not be any contact voltage at the devices. The operating currents, which usually sum up to zero, pass through a summation current transformer whose secondary winding is guided to an overcurrent relay.

Figure 17.97 shows examples of protection techniques.

17.6.5 Energy Storage

Storage Power Stations

Electric energy cannot be stored in its original form. Nevertheless, in electric energy systems the possibility of storing exists by using reservoir power stations. During periods of weak load water is pumped upstream into a reservoir and during peak load periods the potential

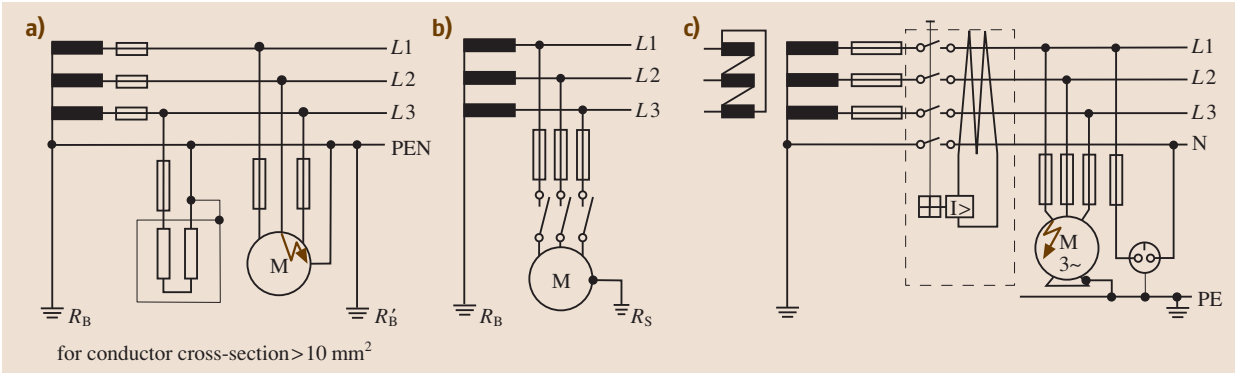


Fig. 17.97a–c Examples of protection techniques: **(a)** grounding without a special grounding conductor; **(b)** protective grounding; **(c)** current-operated earth-leakage circuit breaker, N – neutral conductor, PE – protective grounding conductor (grounded), PEN – zero conductor

energy of this water is used to produce electrical energy again.

For this type of storage system, synchronous machines are used. The synchronous machine is mechanically connected to a water turbine and a pump. A technical variant is the use of only one hydraulic machine, which has to be suitable for turbine as well as pump operation. In this case the operation mode is given by the rotational direction of the machine. The synchronous machine has to be able to start as a motor during pump operation. Reservoir power stations have a total efficiency of energy conversion of about 0.6–0.65.

Batteries

Batteries work according to the basic principle that electrical energy is converted into chemical energy during charging periods, whereas during discharging periods the chemical energy is converted back into electrical energy.

Lead-Acid Storage Battery. These batteries are used for stationary equipment as well as for traction purposes and as starting batteries in vehicles.

Different types of lead-acid battery exist, for example, grid plate accumulators or armor-plate accumulators. The active material is made of lead, lead oxide, and other additives. It is produced in the form of a malleable mass and the solid-lead plates are coated with it. The electrolyte is sulfuric acid with a concentration of about 5 mol/l, and an according density of 1.28 g/cm^3. The concentration decreases during discharging, but should not drop below 1.05 g/cm^3. The total reaction in the cell can be described using the chemical equation

$$\underset{\text{(charging)}}{\mathrm{Pb} + 2\mathrm{H_2SO_4} + \mathrm{PbO_2}} \quad \leftrightarrow \quad \underset{\text{(discharging)}}{2\mathrm{PbSO_4} + 2\mathrm{H_2O}} \,. \tag{17.208}$$

The nominal voltage per cell is defined to be 2.0 V for lead-acid accumulators. The battery's capacity is given in Ah and depends on the type of lead-acid battery used. The charging process can be seen in Fig. 17.98. The electrolyte gas starts to appear at above 2.4 V and charging should be stopped at 2.65 V per cell.

During discharging the minimal voltage should not be below the cutoff voltage of 1.8 V. The useable capacity is a function of the discharge current. Figure 17.98b shows the discharge time as a function of the current for a single cell, using a logarithmic scale. The discharge characteristics of a single cell as a function of the discharge rate with the discharge current as a parameter is presented in Fig. 17.98c. Lead-acid storage batteries are sometimes used as storage devices in the field of electricity supply. Then the battery is used for load leveling, frequency control, provision of instantaneous reserve or voltage control. One example of a 17 MW 14.4 MW h plant was installed and successfully operated in Berlin between 1987 and 1994. A 10 MW 40 MW h plant was used in the Southern California Edison Co. During the

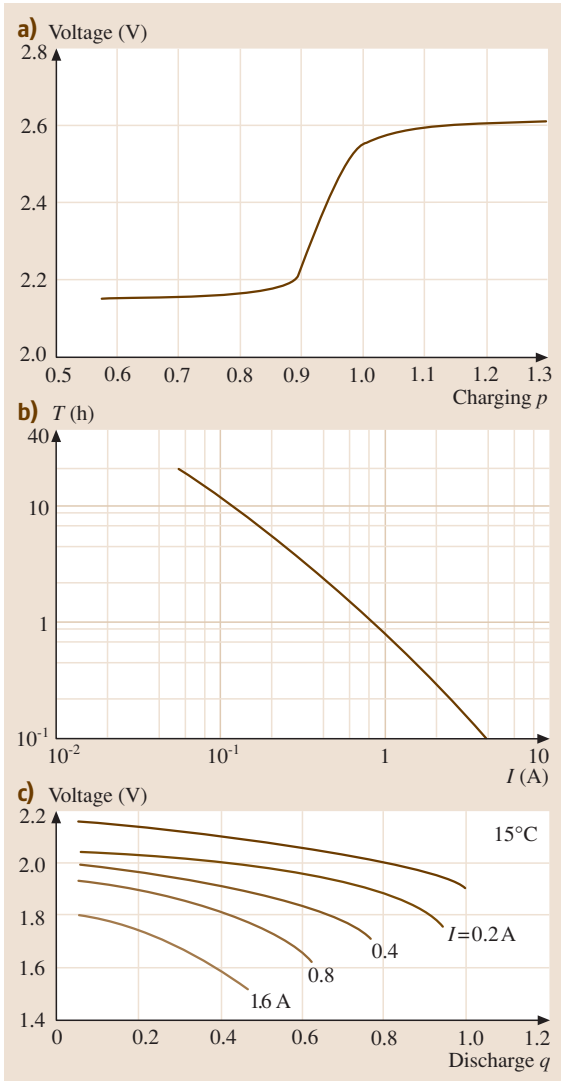


Fig. 17.98a–c Characteristic of lead-acid batteries: **(a)** charge process; **(b)** discharge time of one cell related to discharge current; **(c)** discharge characteristic of a cell at 15 °C

1990s several plants were installed worldwide showing that the economical operation of this system is possible. Developments caused by the liberalization of the energy market may promote this technology.

Other Accumulators. Cadmium and nickel batteries are accumulators with an alkaline electrolyte; they are rechargeable and are produced as button cells with a capacity between 10 mA h and 25 A h. The use of electric drives in vehicles depends considerably on the capability to produce storage batteries with a high energy density. However, the lead-acid accumulator with a specific capacity of 30–40 W h/kg, assuming a 2 h discharging time, cannot entirely fulfil these requirements. A promising development is the sodium/sulphur battery with a ceramic electrolyte and liquid sodium and sulphur. An energy density of about 120 W h/kg is achieved with this technology. The sodium/sulphur battery requires a minimum operating temperature of 285 °C.

Other Energy Storage Methods

Future development is addressing the possibility of storage of magnetic energy in superconducting coils, in which the current can flow nearly without losses. The coupling in and coupling out of electrical energy is realized by an electronic power converter. Other methods involve the unconventional charging processes with switches, which are converted from a superconducting state into a normal state.

A different approach is the use of flywheel storage, which is charged by an electric source and can be discharged depending on the demand for electricity.

17.6.6 Electric Energy from Renewable Energy Sources

Renewable energy is sometimes called *green energy* because it is – in contrast to fossil energy like coal, oil, and gas – not finite. Renewable energy sources are based on regenerating natural primary energy sources (biomass, biogas) or infinite primary energy sources (sun, wind, water). The transformation process into electrical energy has no or only a small effect on the environment and does not contribute to the greenhouse effect and global warming like the conventional burning of fossil sources. The five most common renewable sources – which are finally all solar sources – are hydropower, wind energy, solar energy (thermal as well as photovoltaic), geothermal energy and biomass. Sometimes using fuel cells for the transforming process into electrical energy is also referred to as using a renewable energy source, but this depends on the primary energy – the *fuel* being used. Some of them will now be briefly discussed.

Solar Energy

Solar energy from the sun is the most abundant energy form on the planet. The sun provides energy in two forms – light and heat [17.8]. The thermal energy can be used either for heating (homes, offices, swimming pools etc.) or to power a conventional turbine to produce electricity (a solar thermal power plant). The light energy can be converted into electrical energy directly by using photovoltaic systems.

Photovoltaic Energy Systems. Photovoltaic (PV) energy systems transform light directly into electricity. They use daylight to supply ordinary electrical equipment, for example, household appliances, computers, and lighting.

A PV cell consists of two or more thin layers of semiconducting material, most commonly silicon (for the different types of PV see Table 17.7). When the silicon is exposed to light, electric charge carriers are generated and an electric potential develops between the two layers of the material. To balance this potential a charge transport is necessary, which is conducted via metal contacts on both sides of the cell. If an electrical load is connected to the contacts a direct current (DC) will flow.

To achieve a higher voltage and power level single cells are connected electrically. For mechanical stability and outdoor use these cells are then encapsulated, usually behind glass, to form a module. The connection of several modules is called a panel or string.

The PV module is the principle building block of a PV system and any number of modules can be connected to give the required power with a suitable current and voltage output.

Typical modules have a rated power output of around 75–200 W peak (W_p) each. A typical domestic system of 1.5–2 kW_p may therefore contain about 10–24 modules covering an area of 12–40 m^2. The actual power output is dependent on the technology used and the orientation of the array with respect to the sun. An inverter (see Sect. 17.4) is used to convert the comparatively low DC voltage to a compliant AC voltage.

The solar generators are suitable for autonomous distributed energy supplies as well for feeding into public networks. Figure 17.100 shows an example of a 10 kW_p photovoltaic system.

Table 17.7 Types of PV cells

Type of PV cell	Specification
Monocrystalline silicon cells	• Made from a single cylindrical crystal of silicon • High efficiencies, typically around 15% • Manufacturing process is complicated, resulting in slightly higher costs than other technologies
Multicrystalline silicon cells	• Made from an ingot of melted and recrystallized silicon • Multicrystalline cells are cheaper to produce than monocrystalline cells • Efficiency around 12% • Creating a granular texture
Thick-film silicon	• Another multicrystalline technology • Encapsulated in a transparent insulating polymer with a tempered glass cover • Bound into a strong aluminum frame
Amorphous silicon	• Composed of silicon atoms in a thin homogenous layer • Thinner because of more effective light absorption • Flexible for curved modules • Typical efficiencies of around 6% • Cheaper to produce

PV equipment has no moving parts and as a result requires minimal maintenance. It generates electricity without emitting any greenhouse or other gases, and its operation is silent.

Photovoltaic power is the same technology that supplies certain calculators and watches. It is also used for remote telephones in some regional areas.

Solar Thermal Energy Systems. This type of energy occurs when energy from the sun is absorbed and used to heat liquids such as water or oil. The hot water can either be used directly or via a heat exchanger or to create steam and drive a small turbine to generate electricity.

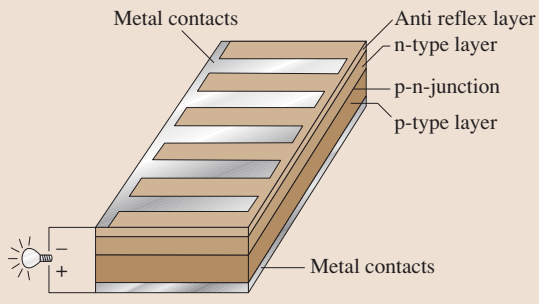


Fig. 17.99 PV cell schematic construction

Advantages of using solar energy:

- Solar energy is infinite and always available.
- It is useful for remote areas that are not connected to the main electricity grid.
- No output of greenhouse gas.

Disadvantages of using solar energy:

- It is weather dependent.
- It is expensive.
- Due to its efficiency it requires a lot of space.

Wind Energy

Wind is moving air that results from different temperature zones created by solar energy. By using wind turbines, kinetic wind energy can be used to generate electrical energy [17.29, 30].

Wind turbines consist of numerous components, including the blades, a shaft, a generator, and a tower. The shaft is connected to the blades, and rotates as the blades are turned by the wind (Fig. 17.101). It can either be connected directly or via a gear box to the generator, which converts the kinetic energy into electrical energy.

The power P of a wind energy converter is proportional to the coating area of the rotor blades A and the cube of the wind speed v^3. The maximal theoretic aerodynamic efficiency of an ideal wind turbine is

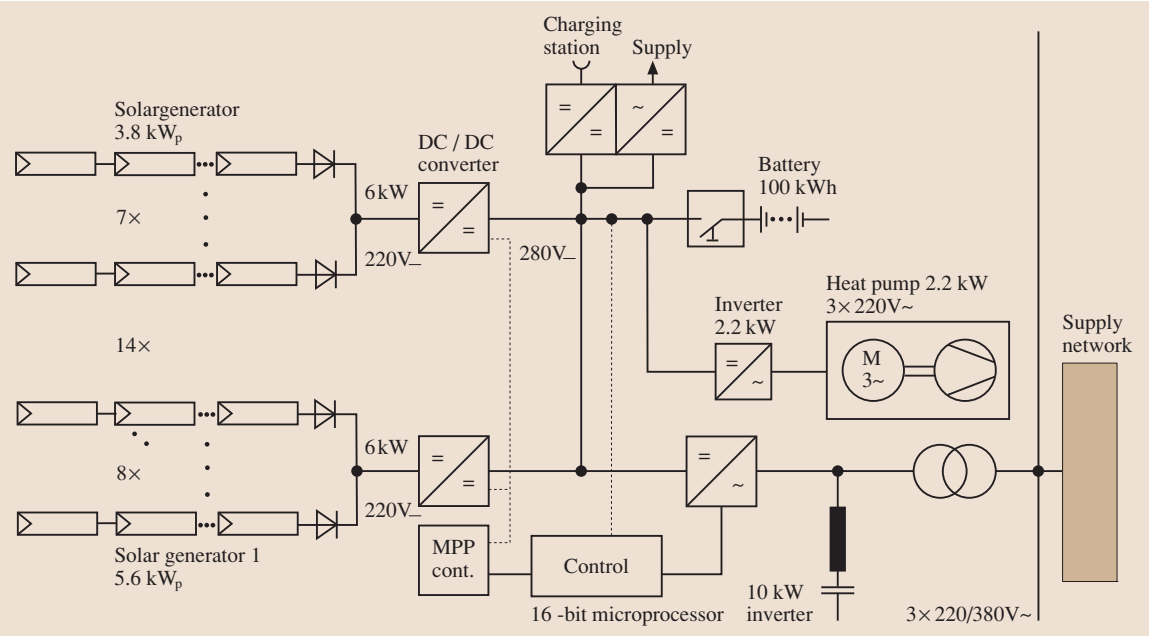


Fig. 17.100 PV system

59.3% [17.31]. Today efficiencies are 38–50% depending on the technology used.

The nominal power of a wind turbine ranges from a few watts up to 7 MW (2008).

Figure 17.103 shows a typical power characteristic as a function of the rotational speed and for different wind speeds. The maximum power point moves to a higher rotation speed with increasing wind speed v. If the generators are connected to a network with constant frequency they have to be adapted to the grid frequency because they cannot work at any wind speed at their maximum power point. This can be solved by applying, for example a link converter between the generator and the three-phase network.

Today four types of machines are used within wind turbines: asynchronous machines, synchronous machines, double-fed induction machines, and full converter machines.

The double-fed induction generator (DFIG) makes it possible to operate the turbine at variable speed. Thereby the utilization even at low wind speeds is enhanced. The present option of choice is based on full AC–DC–AC conversion with a gearless mechanical implementation.

There are various methods for connecting wind turbines to an electrical network:

- A cyclo- or matrix converter
- A converter with a DC link
- A converter with an AC link

The type of connection chosen depends on the kind of wind turbine used and the network parameters (voltage, current, frequency, harmonics, etc.).

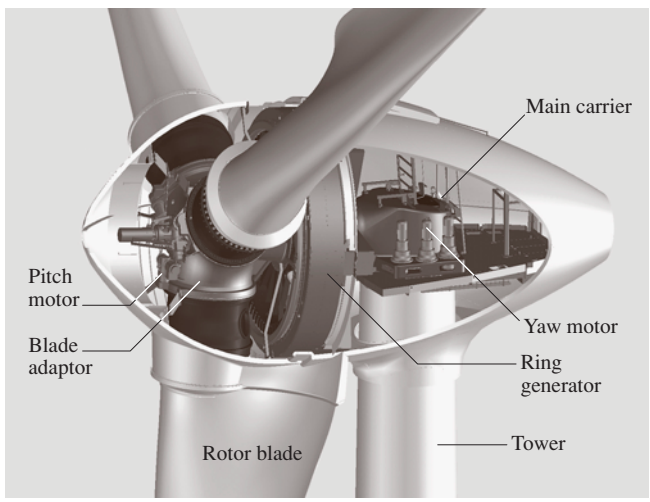


Fig. 17.101 Wind generator (Enercon)

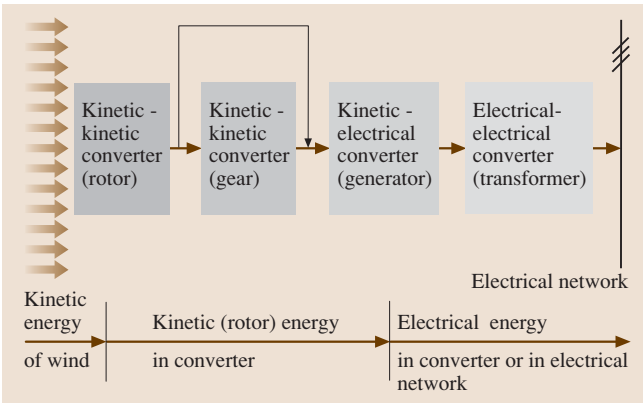


Fig. 17.102 Electrical conversion chain of a wind turbine

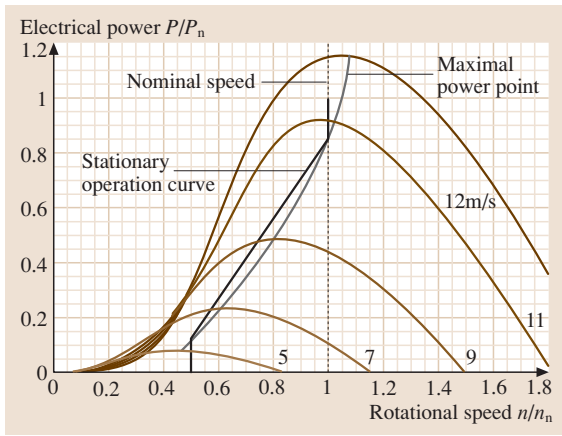


Fig. 17.103 Characteristic of output power against rotational speed and wind speed in per unit (p.u.)

Wind turbines can differ depending on:

- The setting of the rotor axis (horizontal or vertical)
- The number of rotor blades (1–4)
- The tip speed ratio (slow or fast)
- The possibility of power control (pitch or stall)
- The possibility of wind storm security
- The type of generator (synchronous, asynchronous, DFIG or full converter machine),
- The type of network connection (direct or via a DC link)

Advantages of using wind energy:

- Energy from the wind is infinite and always available.
- It is very useful for remote regions that are not connected to the main electricity grid.
- It is environmentally safe.
- Land for wind farms can usually be used for other things as well (farming).

Disadvantages of using energy from the wind:

- Wind speed is variable.
- It creates noise, so turbines should not be built near houses.

Hydropower

The fall and flow of water is part of a continuous natural cycle. The sun draws moisture up from the oceans and rivers, and the moisture then condenses into clouds in the atmosphere. This moisture falls as rain or snow, replenishing the oceans and rivers. Gravity drives the water, turning ist potential energy into kinetic energy, which can be quite high.

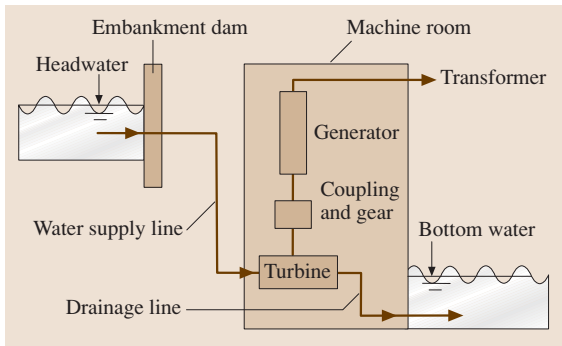


Fig. 17.104 Hydropower station – physical principles

An advantage of hydropower is its operating costs – it is less expensive than mining fossil fuels. Also, unlike other renewable sources like the sun or wind, water can be stored, which makes it interesting for generating electricity.

Rivers, dams, and waterfalls can be used to generate electricity. Today the exploitation of wave energy and tidal power plants are also being developed.

Hydroelectric power stations are built where water is running. The most common location is at dams, where water is stored. To produce hydroelectric power, water from rain or melting snow is collected and stored in a natural or artificial sea. The flow of this water can be controlled by the opening and closing of gates or pipes. The dam wall can also create a high water level, which creates a higher pressure in the pipes to the turbine. A large pipe carries the water from the dam to the turbine. The pressure of the water pushes against the blades and turns the turbines.

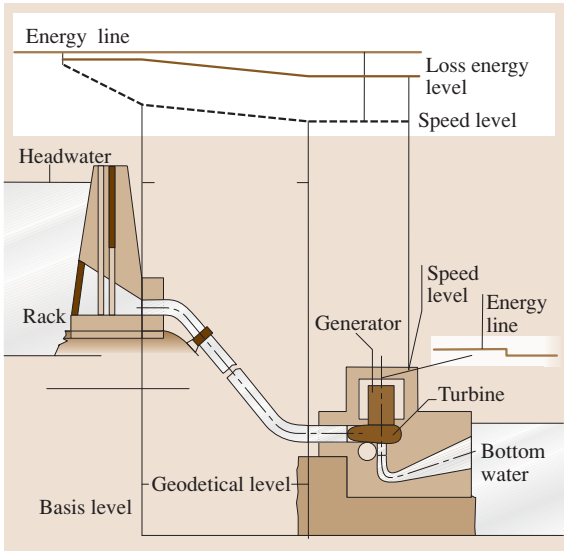


Fig. 17.105 Scheme of a hydropower station

The theoretical power of water is defined as

$$P_{\mathrm{WAth}} = \rho_{\mathrm{WA}} g q_{\mathrm{WA}} (h_{\mathrm{HW}} - h_{\mathrm{BW}}) \,, \qquad (17.209)$$

where ρ_{WA} is the water density, g is the gravitational constant, q_{WA} is the flow rate through hydropower system, h_{HW} is the geodetical level of head water surface, and h_{BW} is the geodetical level of the lower water surface.

The construction of a typical hydropower station is illustrated in Fig. 17.105. It consists of:

- The turbine, a hydraulic machine that transforms the kinetic water energy to rotary motion (it can be a Pelton, Francis, propeller, Kaplan or Straflo turbine).
- Coupling and gear (via direct coupling for small stations or via gears for large units).
- The generator (synchronous or asynchronous generator).
- The transformer (voltage adjustment).
- The embankment dam (responsible for continuous water charge).
- The water supply line (water can flow to the turbine either directly or through a gear canal).

Pumped Storage Systems. Some hydro plants also use pumped storage systems. A pumped storage system operates much as a public fountain does. The same water is used again and again. At a pumped storage hydro plant, flowing water is used to generate electricity and is then stored in a lower pool. Depending on how much electricity during a certain period is needed, the water is pumped back to the upper pool. Pumping water to the upper pool requires electricity, thus this pumping procedure is done during low load periods, when the electricity price is lower than during peak load hours, when selling this energy is beneficial. Pumped storage systems are used when there is a high demand for peak load electricity and the geographical prerequisites are fulfilled as well.

For a long time hydropower has been the leading renewable energy source. Worldwide, hydropower provides more than 90% of all electrical energy produced from renewable sources. All other renewable sources produce only about 10% of the renewable energy (status 2001). In Germany the share of hydropower refering to the electricity consumption was overtaken by wind energy in 2004, while other power systems like e.g. the Norwegian system are mainly based on hydro power.

Advantages of hydropower:

- It is a renewable source
- Hydroelectricity produces no gas emissions or waste.
- It is more reliable than solar and wind power because water can be stored.
- Hydroelectric stations are inexpensive to operate.

Disadvantages of hydropower:

- Large dams take up large areas of land and can cause fish and other animals to relocate.
- The durability of the plant can be affected by a change in water quality.

Fuel Cells as an Alternative Energy Source

In principle, a fuel cell operates like a battery [17.32, 33]. However, unlike a battery, fuel cells do not run down or require recharging. Rather they deliver electricity and heat as long as fuel is supplied. A fuel cell

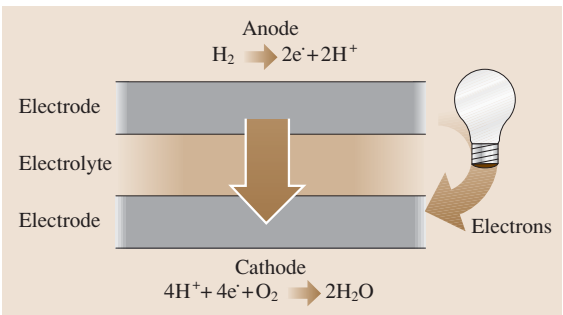


Fig. 17.106 Principles of a fuel cell

consists of two electrodes with an electrolyte between. The most well developed fuel cell is the polymer electrolyte fuel cell (PEMFC) [17.34]. Oxygen passes over one electrode and hydrogen over the other, generating electricity, water, and heat (Fig. 17.106). Hydrogen fuel is fed into the *anode* of the fuel cell. Oxygen (or air) enters the fuel cell through the cathode. With the help of a catalyst, the hydrogen atom splits into a proton and an electron, which take different paths to the cathode. The proton passes through the electrolyte, while the electron creates a separate current that can be utilized before returning to the cathode, to be reunited with the hydrogen and oxygen in a molecule of water. This DC current is transformed into an AC current using a DC–AC converter. Afterwards it will be supplied to the electrical network. For fuel cells the DC–AC converters have to satisfy specific requirements, for example, wide DC voltage range. There are many kinds of DC–AC converters and choosing the suitable one also depends on the circuit and network parameters.

A fuel cell system which includes a *fuel reformer* can utilize the hydrogen from any hydrocarbon or alcohol fuel: natural gas, ethanol, methanol, propane, and even gasoline or diesel. Hydrogen can be extracted from novel feed stocks such as landfill gas or anaerobic digester gas from wastewater treatment plants, from biomass technologies, or from hydrogen compounds containing no carbon, such as ammonia or borohydrides. Hydrogen can also be produced from electricity from conventional, nuclear or renewable sources. Electrolysis uses an electric current to extract hydrogen from water. Fuel cell in combination with solar or wind power, or any renewable source of electricity can contribute to zero-emission energy system that requires no fossil fuels and is not limited by variations of sunlight or wind. The hydrogen from the electrolysis can supply energy for power needs and for transportation.

Fuel cells differ from one another in terms of operation temperature, electrolyte, fuel, kind of application, etc.

Application possibilities suchas the provision of backup power to a grid-connected customer, in case the grid should fail are possible by special configuration. A configuration to provide completely grid-independent power is possible as well, then the grid can be used as a backup system. Modular installation (the installation of several identical units to provide a desired quantity of electricity) provides high reliability, which is crucial to an economy that depends on increasingly sensitive computers, medical equipment, and machines.

Since the fuel cell relies on chemistry and not on combustion, emissions from this type of a system are much smaller than emissions from today's cleanest fuel combustion processes, while also being, fundamentally more efficient than combustion systems:

- Fuel cell power generation systems in operation today achieve 40% fuel-to-electricity efficiency utilizing hydrocarbon fuels.
- Systems fueled by hydrogen consistently provide 50% efficiency. Even more efficient systems are under development.
- In combination with a turbine, electrical efficiencies can exceed 60%.
- If waste heat is put to use for heating and cooling, overall system efficiencies exceed 85%.

17.6.7 Power Quality

Fundamental Terms

The term *power quality* covers a lot of issues which are related to description, analyses, monitoring, and elimination of undesirable phenomena appearing in power systems. Interest in power quality increases due to the changes of the network operation conditions over the last years [17.35].

The significant increase of power demand and the appearance of nonlinearity of loads have resulted in a situation where network elements are overloaded and required to work at their upper load limits. Such properties contribute to unstable network operation,

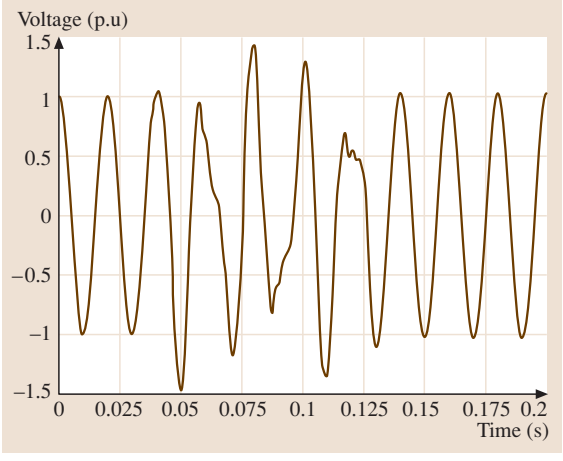


Fig. 17.107 Example of flicker caused by arc furnace operation

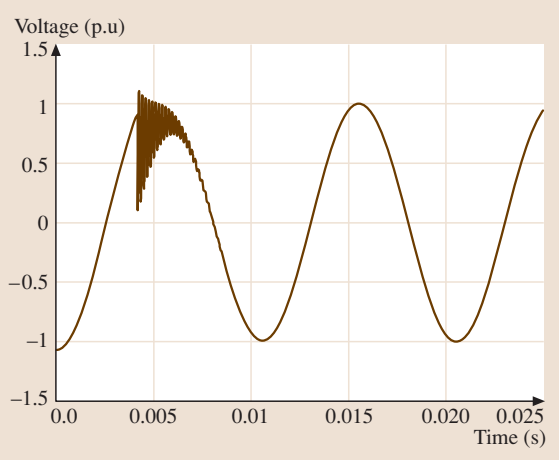


Fig. 17.108 Low-frequency oscillatory transient caused by capacitor bank energization

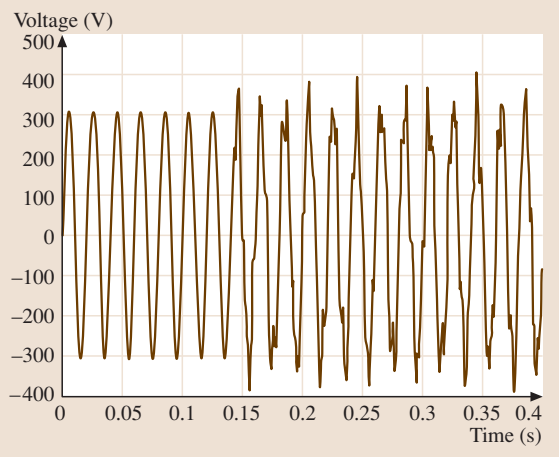


Fig. 17.109 Low-frequency oscillatory transient caused by ferroresonance of an unloaded transformer

known as power quality phenomena (see Figs. 17.107–17.109), which can be classified into the following groups:

- Transients: momentary events. They can be subdivided into two groups: impulsive and oscillatory transients. Impulsive transients are sudden changes in the steady-state condition of voltage or current. The polarity of the impulsive transient is either positive or negative. The oscillatory transient has sudden changes as well, but the polarity has both positive and negative values.
- Long-duration voltage variations: variations of the voltage or current at the power frequency lasting longer than 1 min. These can be divided into several subgroups: overvoltages, which are changes in the RMS AC voltage to more than 110% of the normalized value; and undervoltages, which are decreases of the voltage to less than 90% of the normalized value.
- Short-duration voltage variations: phenomena that last for less than 1 min. The following event types can be distinguished: interruptions appearing when the voltage or current decreases to less than 10%. Sags (dips) are variations at the power frequency in the range between 10–90% in RMS value. Swells are phenomena defined as an increase of the voltage or current in the range 110–180%.
- Blackouts. A black-out is a sustained power outage over a relatively wide area. Usually the pre-stage of a black-out is a so called brown-out. The brown-out leads to a black-out, when the system cannot recover so the outage area grows.
- Voltage imbalances, which arise when the three-phase supply system operates with unsymmetrical voltages.
- Waveform distortions, defined as steady-state deviations from an ideal sine wave. Such events can be divided into the following groups:
  - Harmonics, which are sinusoidal voltages and currents that are integer multiples of the fundamental frequency.
  - Interharmonics, which differ from harmonics in that they have a noninteger relation to the fundamental frequency.
  - Notchings, which are periodic disturbances resulting from the operation of power electronic elements. Noise and DC offset can also be classified into this group of phenomena.

The Power Quality Description

In order to evaluate and analyze power quality, some mathematical definitions need to be introduced. The whole power quality analysis is based mainly on voltage and current studies in the frequency domain. Some specific numerical tools are used, like the discrete Fourier transformation (DFT) and fast Fourier transformation (FFT), which assess the energy quality based on comparative analyses. These analyses can be applied to both impulsive and steady-state phenomena. In order to evaluate the higher harmonic and interharmonic content in the steady state, the total harmonic distortion (THD) is

defined

$$\mathrm{THD} = \frac{\sqrt{\sum_{h>1}^{h_{\max}} G_{\mathrm{h}}^2}}{G_1} , \tag{17.210}$$

where G_{h} is the individual RMS value of the quantity G. It can be seen that THD is referred to as the fundamental harmonic. To measure the distortions produced by nonlinear loads of mainly large power networks the total demand distortion (TDD) is introduced

$$\mathrm{TDD} = \frac{\sqrt{\sum_{h>1}^{h_{\max}} I_{\mathrm{h}}^2}}{I_{\mathrm{L}}} , \tag{17.211}$$

where I_{L} is the peak, or maximum, demand load current at the fundamental frequency. For power quality evaluation the individual harmonic components can be analyzed separately as well. In this case the magnitude of the harmonic cannot exceed the prescribed values that are introduced by standards and norms.

Not only harmonics are evaluated in terms of the power quality. Power analysis is also conducted when the current or voltage occurs as a harmonic signal. In such cases the apparent power S is divided into three components

$$S = \sqrt{P^2 + Q^2 + D^2} , \tag{17.212}$$

where P is the active power at each harmonic, Q is the reactive power at each harmonic, and D is called the distortion power, reflecting the relationship between the different harmonic elements.

The general relationship between the active power (needed to operate equipment) and the apparent power is

$$\mathrm{PF} = \frac{P}{S} , \tag{17.213}$$

the power factor. The best conditions for network operation are obtained when this parameter is equal to 1.

Power Quality Improvement

A lot of methods are used to improve the power quality. These methods are distinguished in terms of phenomena that need to be compensated. Harmonic mitigation, for example, requires the use of filters. Two main groups of filters can be distinguished: passive and active. Passive filters are constructed by a suitable connection of inductances and capacitances. They are designed to reduce individual harmonics (mainly the fifth) or narrow frequency ranges. Over the last few years, active filters have begun to play an important role in the improvement of network operation in terms of harmonics. These elements are based on semiconductor components and a storage system (battery, supercapacitor, superconductor). By applying suitable switching techniques of the valves in active filters the higher harmonics can be limited.

The reactive power in a distribution system has turned out to be an important problem as well. The reactive flow power results in additional losses. To reduce this, capacitor banks are used to compensate the prevailing inductive load.

Furthermore, short circuits are another important issue in network operation. A high short-circuit current can contribute to the destruction of distribution elements (wires, transformers). To avoid short circuits, relaying systems are applied.

Sudden changes in current can induce current in other wires and also cause the aggravation of the power quality. To limit this type of phenomena shielded cables are utilized.

In order to improve the reliability of the network, the designing approach of the power system plays a very important role. Using new numerical techniques, the network can be designed to have more stable operation, and can be monitored to find the distortion source and eliminate it.

17.7 Electric Heating

In resistances electric power is converted into heat [17.1, 36]. This process is described by the simple formula

$$P_{\mathrm{w}} = UI\cos\varphi = I^2 R = U^2/R . \tag{17.214}$$

The different electric heating procedures vary in terms of the kind of resistance and the nature of the energy source and include resistance, arc, inductive, and dielectric heating. For a smelting furnace, the first three principles, shown in Fig. 17.110, are used.

For the techniques illustrated in Fig. 17.110a,e,f, the heated material itself acts as the resistance, and for the examples shown in Fig. 17.110a,d the heating source is arc plasma.

The only practical application of Fig. 17.110c is the arc-heating steel furnace. Most reduction furnaces, for

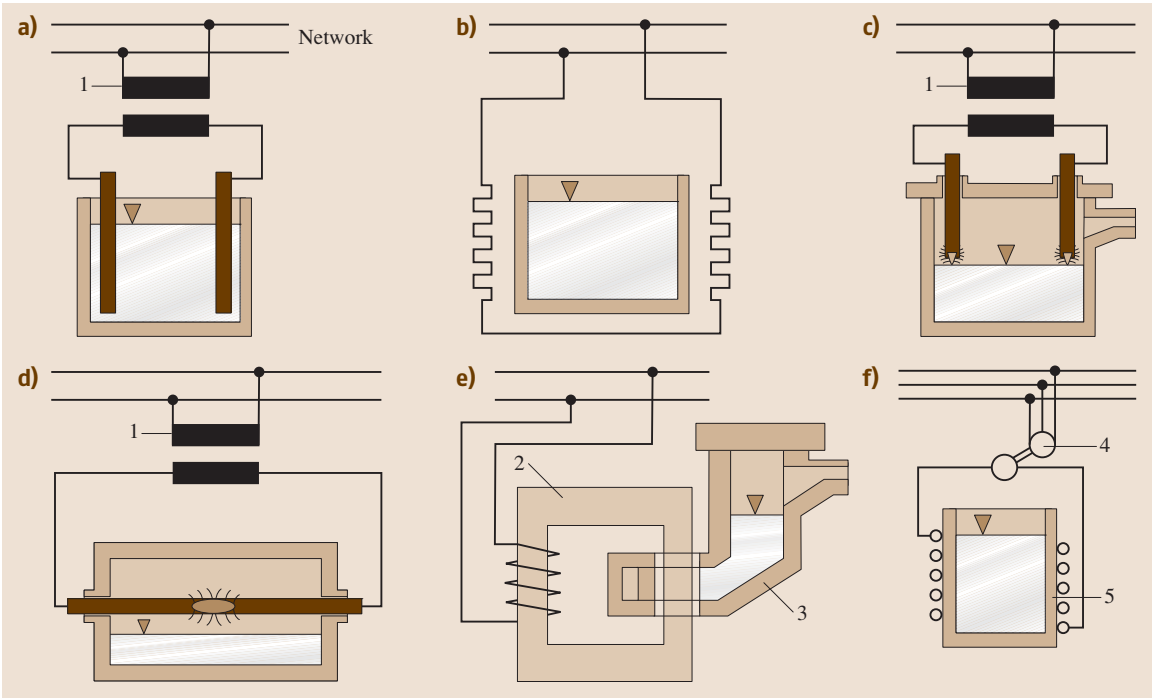


Fig. 17.110a–f Examples of electric heating: **(a)** direct resistance heating; **(b)** indirect resistance heating; **(c)** direct electric arc-heating; **(d)** indirect arc-heating; **(e)** low-frequency induction heating; **(f)** middle-frequency induction heating: 1 – transformer, 2 – iron core, 3 – smelting chamber, 4 – middle-frequency converter, 5 – induction coil with water cooling [17.36]

example, for the production of carbide, ferrosilicon, and aluminum, are hybrid forms of (a) and (c).

17.7.1 Resistance Heating

Figure 17.111 shows examples of direct resistance heating. Such devices are characterized by high current demand at relatively low voltages.

For indirect resistance heating, heating conductors that meet the relevant thermal, chemical, and mechanical demands are needed [17.37]. The heating conductors differ from each other greatly in terms of the temperature dependency of their conductivity. Depending on the temperature range, the following metals can be used: molybdenum, molybdenum compounds, tantalum, and in special cases platinum or ceramic (mainly silicon carbide) and graphite.

Metallic heating conductors are produced in different forms, e.g., wires or bands, whereas the other heating conductor forms are confined to pipe or slat forms. Liquid glass can be used as a heating conductor as well. The heating conductors are mounted on ceramics or other temperature resistant isolators.

17.7.2 Electric Arc Heating

As well as in steel furnaces arcs are also applied in the arc welding process. These arcs can be described as resistances that are dependent not only on their length but also on the current. The relationship between the arc voltage and current is nonlinear (Fig. 17.112).

Arc Furnaces

Large arc furnaces are low resistance loads (a few mΩ and currents of more than 10 000 A at voltages of a few hundred V). The three-phase current that feeds the arc is fed through a graphite electrode. The current is regulated by changing the arc impedance by controlling the distance between the electrodes. In addition, the input voltage at the clamps of the furnaces is changeable to achieve a suitable current level. In order to adjust the voltage, a regulating transformer is used

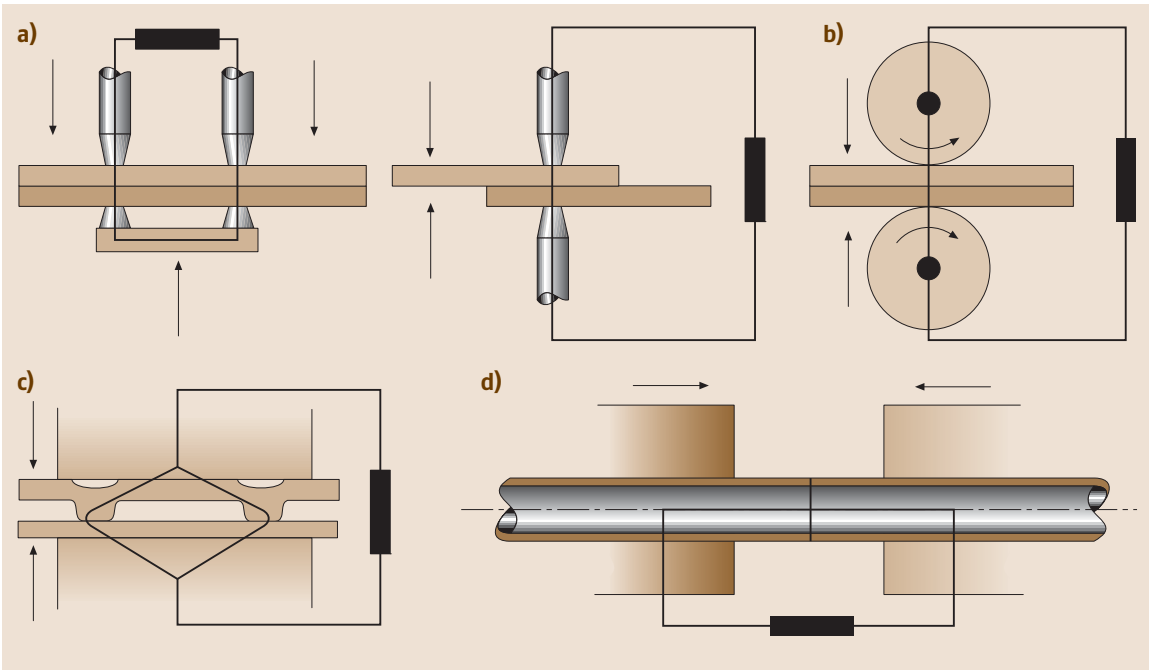


Fig. 17.111a–d Materials and electrode position of resistance welding: **(a)** point welding; **(b)** continues rolling welding; **(c)** projection welding; **(d)** butt welding [17.36]

which has winding taps on the high-voltage side of the transformer, because the high-current (secondary) side consists of only one or a few windings. Thus the furnace voltage is changed through the magnetic flux in the transformer.

The strongly varying magnetic fields on the secondary side evoke high inductive voltage drops that limit the short-circuit current and the maximum voltage on the secondary side. They can also be responsible for an unsymmetrical network load.

The electrodes can be controlled hydraulically or with the help of electric machines. The control variable is the arc impedance, which can only be measured with the help of specific methods; otherwise measurement errors occur due to the magnetic fields generated by the high-current circuit. The voltage v_{0M} in Fig. 17.113 is the error occurring in the normal measurement method.

The inductances L_1, L_2, and L_3 are the real mutual inductances of the high-current loop, namely

$$L_1 = M_{12,13}\frac{\mathrm{d}i_1}{\mathrm{d}t} + M_{23,21}\frac{\mathrm{d}i_2}{\mathrm{d}t} + M_{31,32}\frac{\mathrm{d}i_3}{\mathrm{d}t} \,. \tag{17.215}$$

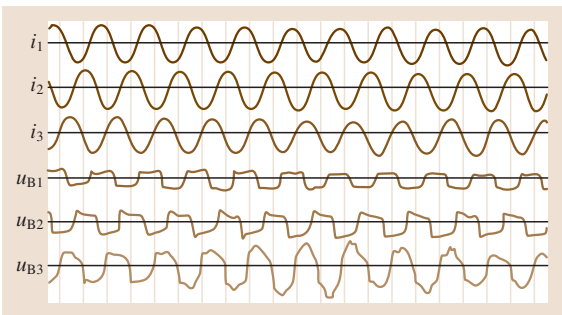


Fig. 17.112 Current and arc voltage of an arc furnace

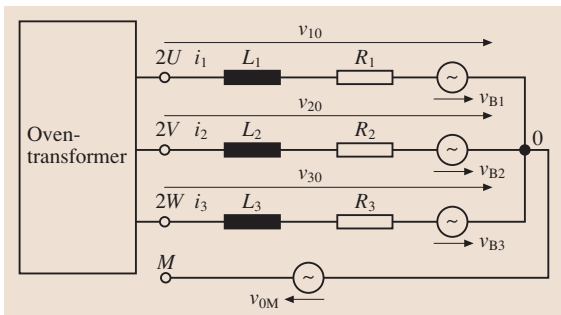


Fig. 17.113 Equivalent circuit for an arc furnace

The error voltage in the star point of the circuit results from the expression

$$V_{0M} = M_{2M,M3}\frac{di_1}{dt} + M_{3M,M1}\frac{di_2}{dt} + M_{1M,M2}\frac{di_3}{dt}\,. \tag{17.216}$$

where v_{B1}, v_{B2}, and v_{B3} are the arc voltages and R_1, R_2, and R_3 are the resistances of the high-current circuit (Fig. 17.113).

Arc Welding

Arc welding requires voltages of up to 40 V, which are not dangerous values for manual welding. The DC or AC voltage source is connected to the workpiece with one pole and to the electrode (welding rod) with the other pole (for DC sources mainly the negative pole). In order to melt the material properly, a suitable arc feeding current and voltage level must be maintained. Therefore, the voltage source must match the characteristics of the arc.

17.7.3 Induction Heating

For inductive heating, the transformer principle is employed to transfer the power to the workpiece, which has the function of the short circuit secondary winding. In crucible furnaces (Fig. 17.110f) the secondary winding has a massive cylindrical form, in contrast to the ring form in trench furnaces (Fig. 17.110e). Therefore an iron core is not used in a crucible furnace.

Due to the iron losses in trench furnaces, they are only used for frequencies of 50–60 Hz. The crucible furnace, on the other hand, can operate in a frequency range up to 10 kHz.

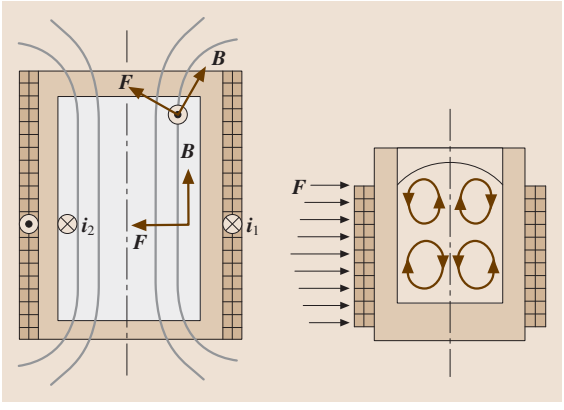


Fig. 17.114 Field concentration in a crucible furnace

Skin Effect, Depth of Penetration

The skin effect of the eddy current (secondary current) causes the current density to be higher in places that lie near to the induction coil. The depth of penetration δ is the depth x measured from the surface nearest to the primary coil for which the current density S has a value of S_0/e, where S_0 is the current density at the surface and e is the Euler number (≈ 2.716).

The following equation is valid for the current density

$$S(x,t) = \hat{S}_0 \exp(-x/\delta)\cos(\omega t - x/\delta) \tag{17.217}$$

and, for the depth of penetration

$$\delta = \sqrt{\frac{2}{\omega\mu\kappa}} = 503\sqrt{m/\Omega s}\,\frac{1}{\sqrt{f\mu_r\kappa}}\,. \tag{17.218}$$

δ is thus inversely proportional to the root of the frequency, the permeability, and the conductance of the

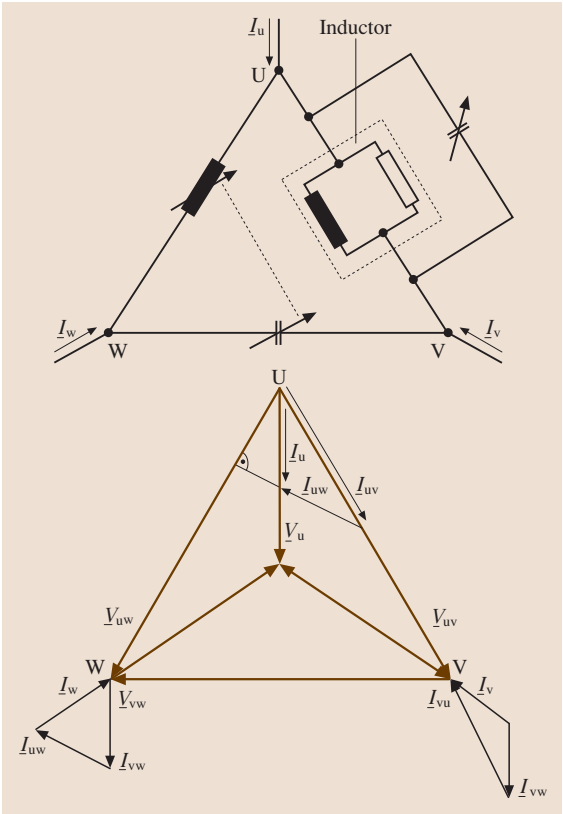


Fig. 17.115 Method to compensate the unsymmetrical and reactive power load of an inductive heating facility

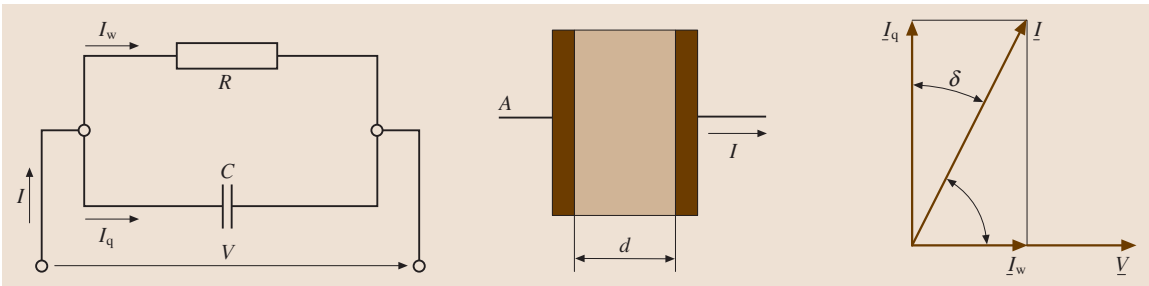


Fig. 17.116 Equivalent circuit and phasor diagram for a real capacitor for dielectric heating

material. Using higher frequencies the current density and therefore the heating can be concentrated to the surface of the materials.

Surface Bulging and Melt Circulation in the Heating Material

Currents in the inductance coil and the materials produce repulsive forces. In the melted materials they are directed towards the axis of the cylinder. This causes the static pressure to be higher at the cylinder axis than at its perimeter. As a result the upper surface of the melted material forms a bulge Fig. 17.114. Because the magnetic field contains radial components at the ends of the cylinder, torsional moments arise in the melted material which cause vorticity of the material. The bulging end and torsional moments are approximately proportional to $\sqrt{1/f}$ so they can be adapted to the requirements of the process by changing the frequency.

High-Frequency Induction Surface Heating

By using high frequencies with suitably formed induction coils, selective surface areas of the working piece can be heated to achieve surface improvement. If the input power has a high level, then the surface can be heated faster, while the heat has no time to be conducted to the inner material. Thus the depth of heating is controlled by the input power and as well by the frequency.

Electric Power Supply

The facilities of induction heating are generally single-phase loads mainly working at a frequency higher than the network frequency. Figure 17.115 shows a common method to turn the load into a symmetrical load, free of reactive power. The current in the coil is compensated by an adjustable capacitor to achieve $\cos\varphi = 1$.

If the reactance of the capacitances and inductances, which are connected to the third phase of the three-phase system, are regulated to the value of $\sqrt{3}$ of the value from the heating facility's resistance (which depends on the amount and type of the workpiece) than from the network point of view the load is symmetrical. For applications with intermediate frequencies ($< 30\,\text{kHz}$) and power of up to several megawatts thyristor converters are used. At lower power, IGBT converters that operate with frequencies of up to 500 kHz are employed. In both cases the converters have resonant circuit topologies.

17.7.4 Dielectric Heating

A high-frequency alternating electric field causes losses in non-conducting materials due to changes in the position of the molecular dipoles. In this process the phase angle of the input current is $\varphi < 90°$ with respect to the given voltages (Fig. 17.116). The angle δ is called the loss angle and is nearly independent of the frequency.

The following expression describes the specific losses and heating power

$$P'_w = E^2 \omega \varepsilon_0 \varepsilon_r \tan\delta\,,$$

$$\text{where } \tan\delta = \frac{I_w}{I_g} = \frac{P_w}{P_g}\,. \qquad (17.219)$$

The electric field intensity is limited by the breakdown field strength of the dielectric. Further increases of the power density P'_w are only possible by changing the frequency. The value of $\tan\delta$ depends on the temperature and humidity of the material and can often be used for self-regulation of the power P'_w. Because this technology is often used with multilayered material the orientation of the boundary layer can be lengthwise or transverse to the electric field. Which orientation is best depends on the required process parameters. Air in the field volume affects the process adversely.

Dielectric heating is very often applied for welding of plastic foils, particularly in the form of joints, for example, for upholstery or car interior linings.

The suitable power density is only achieved at high frequencies. The following frequencies are authorized for this process: 13.6, 27.12, 40.68, and 433.92 MHz. For the heating process in a microwave radiation field the frequency of 2450 MHz is permitted. When using other frequencies, shielding must be applied in order to achieve EMC compatibility. The mentioned frequencies are produced by electron valve generators.

References

17.1 W. Hoffmann, M. Stiebler: Elektrotechnik. In: *Dubbel*, ed. by K.-H. Grote, J. Feldhusen (Springer, Berlin, Heidelberg 2007), in German

17.2 K. Küpfmüller, W. Matthis, A. Reibiger: *Theoretische Elektrotechnik*, 17. Aufl. (Springer, Berlin 2006), in German

17.3 H. Fischer, H. Hofmann, J. Spindler: *Werkstoffe in der Elektrotechnik*, 5. Aufl. (Hanser, München 2002), in German

17.4 G. Lehner: *Elektromagnetische Feldtheorie*, 5. Aufl. (Springer, Berlin 2006), in German

17.5 W.-K. Chen: *The Circuits and Filters Handbook*, 2nd edn. (CRC, Boca Raton 2003)

17.6 E. Ivers Tiffee, W. von Müch: *Werkstoffe der Elektrotechnik*, 9. Aufl. (Stuttgart, Teubner 2004), in German

17.7 D.G. Fink: *Standard Handbook for Electrical Engineers*, 14th edn. (McGraw-Hill, New York 2000)

17.8 T. Markvart: *Solar Electricity*, 2nd edn. (Wiley, Chichester 2000)

17.9 R.M. DelVecchio: *Transformer Design Principles* (Gordon Breach, Amsterdam 2001)

17.10 J.H. Harlow: *Electric power transformer engineering* (CRC, Boca Raton 2004)

17.11 IEC: *IEC60747-9: Semiconductor devices – Discrete devices – Part 9 – Insulated-gate bipolar transistors IGBTs*, 1st edn. (International Electrotechnical Commission, Geneva 2001)

17.12 E.J. Rothwell: *Electromagnetics* (CRC, Boca Raton 2001)

17.13 R.H. Bishop: *The Mechatronics Handbook* (CRC, Boca Raton 2002)

17.14 G. Holmes, T. Lipo: *Pulse Width Modulation for Power Converters* (IEEE Press, Piscataway 2003)

17.15 IEC: *IEC61000-3-2: Electromagnetic Compatibility (EMC) – Part 3–2 – Limits – Limits for harmonic current emissions – equipment input current ≤ 16 A per phase* (International Electrotechnical Commission, Geneva 2005)

17.16 J. Baliga: *Modern Power Devices* (Wiley, New York 1987)

17.17 IEC: *IEC60747-1: Semiconductor devices – Part 1 – General*, 2nd edn. (International Electrotechnical Commission, Geneva 2006)

17.18 IEC: *IEC60747-2: Semiconductor devices – Discrete devices and integrated circuits – Part 2 – Rectifier diodes*, 2nd edn. (International Electrotechnical Commission, Geneva 2000)

17.19 IEC: *IEC60747-2-1: Semiconductor devices – Discrete devices – Part 2 – Rectifier diodes – Section 1 – blank detail specification for rectifier diodes (including avalanche rectifier diodes), ambient and case-rated, up to 100 A*, 1st edn. (International Electrotechnical Commission, Geneva 1989)

17.20 IEC: *IEC60747-2-2: Semiconductor devices – Discrete devices – Part 2 – Rectifier diodes – Section 2 – blank detail specification for rectifier diodes (including avalanche rectifier diodes), ambient and case-rated, for currents greater than 100 A*, 1st edn. (International Electrotechnical Commission, Geneva 1993)

17.21 IEC: *IEC60747-6: Semiconductor devices – Part 6: Thyristors*, 2nd edn. (International Electrotechnical Commission, Geneva 2000)

17.22 IEC: *IEC60747-8-4: Discrete semiconductor devices – Part 8-4 – Metal-oxide-semiconductor field-effect transistors MOSFETs for power switching applications*, 1st edn. (International Electrotechnical Commission, Geneva 2004)

17.23 IEC: *IEC60747-15: Discrete semiconductor devices – Part 15 – Isolated power semiconductor devices*, 1st edn. (International Electrotechnical Commission, Geneva 2003)

17.24 R.W. de Doncker, J.P. Lyons: *The auxiliary resonant commutated pole converter, Proc. IEEE-IAS Conference* (IEEE, Piscataway 1990) pp. 1228–1235

17.25 U. Riefenstahl: *Elektrische Antriebstechnik*, 2. Aufl. (Stuttgart, Teubner 2006), in German

17.26 J. Vogel: *Elektrische Antriebstechnik*, 6. Aufl. (Hüthig, Heidelberg 1998), in German

17.27 W. Leonhard: *Control of Electrical Drives* (Springer, Berlin, Heidelberg 2001)

17.28 J.N. Chiasson: *Modeling and High-Performance Control of Electric Machines* (Wiley, New York 2005)

17.29 S. Heier: *Windkraftanlagen*, 4. Aufl. (Stuttgart, Teubner 2005), in German

17.30 S. Mathew: *Wind Energy – Fundamentals, Recourse Analysis and Economics* (Springer, Berlin, Heidelberg 2006)

17.31 E. Hau: *Wind Turbines, Fundamentals, Application, Economics* (Springer, Berlin, Heidelberg 2006)

17.32 R. O'Hayre: *Fuel Cell Fundamentals* (Wiley, New York 2006)

17.33 G. Hoogers: *Fuel Cell Technology Handbook* (CRC, Boca Raton 2003)

17.34 N. Sammes: *Fuel Cell Technology* (Springer, Berlin, Heidelberg 2006)

17.35 G.J. Wakileh: *Power Systems harmonics – Fundamentals, Analysis and Filter Design* (Springer, Berlin, Heidelberg 2001)

17.36 M. Rudolph, H. Schaefer: *Elektrothermische Verfahren*, (Berlin, Springer 1989), in German

17.37 C.J. Erickson: *Handbook of Electrical Heating for Industry* (IEEE, New York 1995)

18. General Tables

Stanley Baksi

Table 18.1 Basic dimensions, their symbols, and physical significance

Basic dimension	Symbol	Physical entity
Meter	m	Length
Kilogram	kg	Mass
Second	s	Time
Ampère	A	Electric current
Kelvin	K	Thermodynamic temperature, temperature difference
Mol	mol	Quantity
Candela	cd	Light intesity

Table 18.3 Prefixes for units

Power of ten	Prefix	Symbol
10^{18}	Exa	E
10^{15}	Peta	P
10^{12}	Tera	T
10^{9}	Giga	G
10^{6}	Mega	M
10^{3}	Kilo	k
10^{2}	Hekto	h
10	Deka	da
10^{-1}	Deci	d
10^{-2}	Centi	c
10^{-3}	Milli	m
10^{-6}	Micro	μ
10^{-9}	Nano	n
10^{-12}	Pico	p
10^{-15}	Femto	f
10^{-18}	Atto	a

Table 18.4 Units not defined in the Système International d'Unités (SI) system but commonly used

Characteristics of units	Example
Generally used units	Liter (l), hour (h), degree (deg, °)
Units with limited use	Electron volt (eV)

Table 18.5 Conversion values to calculate from m–kp–s in the SI system

1 kp ≈ 1 daN	1 at ≈ 1 bar	1 kp m ≈ 1 daJ
1 kp/cm ≈ 1 N/mm	1 PS ≈ 0.75 kW	
1 mm water column ≈ 0.1 mbar	1 kcal ≈ 4.2 kJ	

Table 18.6 Names and abbreviations of English units

atm	Atmosphere
bbl	Barrel
btu	British thermal unit
bu	Bushel
cwt	Hundred weight
cal	Calorie
deg F	Degree Fahrenheit
ft	Foot
gal	Gallon
hp	Horsepower
in	Inch
lb	Pound
lbf	Pound force
lntn	Long ton
mi	Mile
pdl	Poundel
shtn	Short ton
yd	Yard
UK	United Kingdom
US	United States of America

in/s ≡ inch per second; in^2 ≡ square inch; in^3 ≡ cubic inch

Table 18.9 Nomenclature for astronomical prefixes

10^{6}	Million	Mega	M
10^{9}	Billion	Giga	G
10^{12}	Trillion	Tera	T
10^{15}	Quadrillion	Peta	P
10^{18}	Quintillion	Exa	E
10^{21}	Sextillion	Zetta	Z
10^{24}	Septillion	Yotta	Y
10^{27}	Octillion	–	–

Table 18.2 Types of units and their explanation

Derived SI – units characteristics	Development of symbols for derived SI units description	Example
Units (combined) without special symbols	These symbols are developed by combining the basic units of their constituting elements, for example, units of area, volume, velocity, etc.	m^2, m^3, m/s
Units (combined) with special symbols	–	Newton (N), Pascal (Pa), Joule (J), Watt (W), Ohm (Ω)
Units (combined) with mixed symbols	These symbols are developed by combining the specific units of their constituting elements and some basic units of measurement. If necessary, a combination of symbols which describe the properties of the measurement	Newtonmeter (N m), Pascalsecond (Pa s)

Table 18.7 Roman counting system

I ≡ 1 V ≡ 5 X ≡ 10 L ≡ 50 C ≡ 100 D ≡ 500 M ≡ 1000					
1	I	10	X	100	C
2	II	20	XX	200	CC
3	III	30	XXX	300	CCC
4	IV	40	XL	400	CD
5	V	50	L	500	D
6	VI	60	LX	600	DC
7	VII	70	LXX	700	DCC
S	VIII	80	LXXX	800	DCCC
9	IX	90	XC	900	CM
Writing direction is from left to right, and the individual values are added to get the value. The lesser values always come after the bigger values. If otherwise, the lesser values have to be subtracted from the bigger values.					
V, L, D			can be written only once.		
I, X, C			can be written up to three times.		
Examples					
1496			MCDXCVI		
673			DCLXXIII		
1891			MDCCCXCI		
1981			MCMLXXXI		

Table 18.12 Commonly used units in thermodynamic calculations

Unit[a]	Symbol	Physical relation	Relation to basic units
Kelvin	K	Thermodynamic temperature, temperature potential	–
Meter square per second	m^2/s	Conductivity	–
Joule	J	Heat energy	$1\,J = 1\,kg\,m^2/s^2$
Watt	W	Heat power	$1\,W = 1\,kg\,m^2/s^3$
Joule per kilogram	J/kg	Specific inner energy	$1\,J/kg = 1\,m^2/s^2$
Joule per Kelvin	J/K	Heat capacity	$1\,J/K = 1\,m^2/(s^2K)$
Joule per kilogram and Kelvin	J/(kg K)	Specific heat capacity	$1\,J/(kg\,K) = 1\,m^2/(s^2K)$
Watt per meter square	W/m^2	Heat flux density	$1\,W/m^2 = 1\,kg/s^3$
Watt per meter square and Kelvin	$W/(m^2K)$	Heat transfer coefficient	$1\,W/(m^2K) = 1\,kg/(s^3K)$
Watt per meter and Kelvin	W/(m K)	Heat conductance	$1\,W/(m\,K) = 1\,kg\,m/(s^3K)$
Kelvin per watt	K/W	Thermal resistance	$1\,K/W = 1\,K\,s^3/(kg\,m^2)$
Degree Celsius	°C	Temperature	1 °C = 1 K

[a] Both SI and other unit systems (commonly used ones) are presented

Table 18.8 Conversion of important units from the foot–pound–second (fps) system to the SI system

	fps (foot-pound-second)	SI (m-kg-s)
Length	1 ft = yd = 12 in	1 ft = 0.3048 m; 1 mi = 1609.34 m
Area	1 ft^2 = 144 in^2	1 ft^2 = 0.092903 m^2
Volume	1 ft^3 = 1728 in^3 = 6.22882 gal (UK) 1 gal (US) = 0.83268 gal (UK)	1 ft^3 = 0.0283169 m^3 1 bu (US) = 35.23931; 1 bbl (US) = 115.6271
Velocity	1 ft/s 1 knot = 1.150785 mi/h = 1.6877 ft/s	1 ft/s = 0.3048 m/s
Acceleration	1 ft/s^2	1 ft/s^2 = 0.3048 m/s^2
Mass	1 lb = cwt/112; 1 shtn = 2000 lb 1 slug = 32.174 lb; 1 lntn = 2240 lb	1 lb = 0.453592 kg 1 slug = 14.5939 kg
Force	1 lbf 1 pdl = 0.031081 lbf	1 lbf = 4.44822 N 1 pdl = 0.138255 N
Work	1 ft lb = 0.323832 cal_{IT} 1 btu = 252 cal_{IT} = 778.21 ft lb	1 ft lb = 1.35582 J 1 btu = 1.05506 kJ
Pressure	1 lb/ft^2 = 6.9444 × 10^{-3} lb/in^2 1 lb/in^2 = 0.068046 atm 1 atm = 29.92 in Hg = 33.90 ft water	1 lb/ft^2 = 47.88 N/m^2 1 lb/in^2 = 6894.76 N/m^2 1 atm = 1.01325 bar
Density	1 lb/ft^3 = 5.78704 × 10^{-4} lb/in^3 1 lb/gal = 6.22882 lb/ft^3	1 lb/ft^3 = 16.0185 kg/m^3 1 lb/gal = 99.7633 kg/m^3
Temperature	32°F = 0 °C, 212°F = 100 °C	1°F = 0.5556 °C
Power	1 ft lb/s = 1.8182 × 10^3 hp = 1.28505 × 10^{-3} btu/s	1 ft lb/s = 1.35582 W
Specific heat capacity	1 btu/(lb°F)	1 btu/(lb°F) = 4.1868 kJ/(kg K)
Thermal conductivity	1 btu/(ft h°F)	1 btu/(ft h°F) = 1.7306 W/(m K)
Heat transfer coefficient	1 btu/(ft^2 h°F)	1 btu/(ft^2 h°F) = 5.6778 W/(m^2 K)
Viscosity, kinematic	1 ft^2/s	1 ft^2/s = 0.092903 m^2/s
Viscosity, dynamic	1 lb/(ft s)	1 lb/(ft s) = 1.48816 kg/(m s)

Table 18.11 Units commonly used for calculation in mechanics

Unit[a]	Symbol	Physical meaning	Relation to basic units
Kilogram	kg	Mass	–
Kilogram per second	kg/s	Mass flow	–
Kilogram meter square	kg m^2	Mass moment	–
Kilogram per cubic meter	kg/m^3	Density	–
Cubic meter per kilogram	m^3/kg	Specific volume	–
Square meter per second	m^2/s	Kinematic viscosity	–
Newton	N	Force	1 N = 1 kg m/s^2
Pascal	Pa	Pressure	1 Pa = 1 kg/(m s^2)
Joule	J	Work, energy	1 J = 1 kg m^2/s^2
Watt	W	Power	1 W = 1 kg m^2/s^3
Newtonmeter	N m	Moment	1 J Nm = 1 kg m^2/s^2
Newton per square meter	N/m^2	Pressure	1 N/m^2 = 1 kg/(m s^2)
Pascalsecond	Pa s	Dynamic viscosity	1 Pa s = 1 kg/(m s)
Joule per cubic meter	J/m^3	Energy density	1 J/m^3 = 1 kg/(m s^2)
Ton	t	Mass	1 t = 1000 kg
Gram	g	Mass	1 g = 1/1000 kg

[a] Both SI and other unit systems (commonly used ones) are presented

Table 18.10 Space and time units

Unit[a]	Symbol	Physical relation and technical value	Description through the basic units
Meter	m	Length	–
Second	s	Time	–
Square meter	m^2	Surface	–
Cubic meter	m^3	Volume	–
Meter per second	m/s	Velocity	–
Meter per second square	m/s^2	Acceleration	–
Meter cube per second	m^3/s	Volume flow	–
Radian	rad	Angle	$1\,\text{rad} = 1\,\text{m/m}$
Steradian	sr	Three-dimensional angle	$1\,\text{sr} = 1\,\text{m}^2/\text{m}^2$
Hertz	Hz	Frequency	$1\,\text{Hz} = 1\,/\text{s}$
Radian per second	rad/s	Angular velocity	–
Radian per second squared	rad/s^2	Angular acceleration	–
Liter	l	Volume	$1\,\text{l} = 10^{-3}\,\text{m}^3$
Degree	°	Angle	$1° = \pi/180\,\text{rad}$
Minute	′	Angle	$1' = \pi/(180\times60)\,\text{rad}$
Second	″	Angle	$1'' = \pi/(180\times60\times60)\,\text{rad}$
Minute	min	Time	$1\,\text{min} = 60\,\text{s}$
Hour	h	Time	$1\,\text{h} = 60\,\text{min} = 3600\,\text{s}$
1 per second	1/s	Frequency	$1/\text{min} = (1/60)\,1/\text{s}$
			$1/\text{h} = (1/60)\,1/\text{min} = (1/60)^2\,1/\text{s}$

[a] Both SI and other unit systems (commonly used ones) are presented

Table 18.13 Commonly used units in electric current calculations

Unit[a]	Symbol	Physical relation	Relation to basic units
Ampère	A	Electric current	–
Ampère per square meter	A/m^2	Electric current density	–
Ampère per meter	A/m	Electric current distribution	–
Coulomb	C	Electric charge	$1\,\text{C} = 1\,\text{A s}$
Watt	W	Electric power	$1\,\text{W} = 1\,\text{kg m}^2/\text{s}^3$
Volt	V	Electric potential	$1\,\text{V} = 1\,\text{kg m}^2/(\text{A s}^3)$
Farad	F	Electric capacity	$1\,\text{F} = 1\,\text{A}^2\text{s}^4/(\text{kg m}^2)$
Ohm	Ω	Electric resistance	$1\,\Omega = 1\,\text{kg m}^2/(\text{A}^2\text{s}^3)$
Siemens	S	Electric conductance	$1\,\text{S} = 1\,\text{A}^2\text{s}^3/(\text{kg m}^2)$
Coulomb per square meter	C/m^2	Electric flux density	$1\,\text{C/m}^2 = 1\,\text{A s/m}^2$
Volt per meter	V/m	Electric field intensity	$1\,\text{V/m} = 1\,\text{kg m}^3/(\text{A s}^3)$
Farad per meter	F/m	Dielectric constant, electric field constant	$1\,\text{F/m} = 1\,\text{A}^2\text{s}^4/(\text{kg m}^3)$
Ohmmeter	Ω m	Specific electric resistance	$1\,\Omega\,\text{m} = 1\,\text{kg m}^3/(\text{A}^2\text{s}^3)$
Siemens per meter	S/m	Specific electric conductivity	$1\,\text{S/m} = 1\,\text{A}^2\text{s}^3/(\text{kg m}^3)$

[a] Both SI and other unit systems (commonly used ones) are presented

Table 18.14 Commonly used units in magnetic calculations

Unit[a]	Symbol	Physical relation	Relation to basic units
Ampère	A	Magnetic potential	–
Ampère per meter	A/m	Magnetic intensity	–
Weber	Wb	Magnetic flux	$1\,\mathrm{Wb} = 1\,\mathrm{kg\,m^2/(A\,s^2)}$
Tesla	T	Magnetic induction	$1\,\mathrm{T} = 1\,\mathrm{kg/(A\,s^2)}$
Henry	H	Inductivity, magnetic conductance	$1\,\mathrm{H} = 1\,\mathrm{kg\,m^2/(A^2s^2)}$
Henry per meter	H/m	Permeability, magnetic field constant	$1\,\mathrm{H/m} = 1\,\mathrm{kg\,m/(A^2s^2)}$
1/Henry	1/H	Magnetic resistance	$1\,/\mathrm{H} = 1\,\mathrm{A^2s^2/(kg\,m^2)}$

[a] Both SI and other unit systems (commonly used ones) are presented

Table 18.15 Commonly used units in luminosity calculations

Unit[a]	Symbol	Physical relation	Relation to basic units
Candela	cd	Luminosity	–
Candela per square meter	$\mathrm{cd/m^2}$	Light density	–
Lumen	lm	Luminous flux	$1\,\mathrm{lm} = 1\,\mathrm{cd\,sr}$
Lux	lx	Illumination	$1\,\mathrm{lx} = 1\,\mathrm{cd\,sr/m^2}$
Lumen second	lm s	Luminous energy	$1\,\mathrm{lm\,s} = 1\,\mathrm{cd\,sr\,s}$
Lux second	lx s	Exposure	$1\,\mathrm{lx\,s} = 1\,\mathrm{cd\,sr\,s/m^2}$

[a] Both SI and other unit systems (commonly used ones) are presented

Table 18.16 Physical constants

Gravitational constant	G	$6.672 \times 10^{-11}\,\mathrm{N\,m^2/kg^2}$
Acceleration due to gravity	g_n	$9.8067\,\mathrm{m/s^2}$
Gas constant	R	$8314.41\,\mathrm{J/(kmol\,K)}$
Molar volume	V_m	$22.414\,\mathrm{m^3/kmol}$ at 1.01325 bar, 0 °C
Avogadro constant	N_A	$6.0221 \times 10^{26}\,\mathrm{kmol^{-1}}$
Loschmidt constant	N_L	$2.6868 \times 10^{25}\,\mathrm{m^{-3}}$
Boltzmann constant	k_B	$1.3807 \times 10^{-23}\,\mathrm{J/K}$
Electric field constant	ε_0	$8.8542 \times 10^{-12}\,\mathrm{F/m}$
Magnetic field constant	μ_0	$1.2566 \times 10^{-6}\,\mathrm{H/m}$
Electric charge	e	$1.6022 \times 10^{-19}\,\mathrm{C}$
Faraday constant	F	$9.6485 \times 10^{7}\,\mathrm{C/kmol}$
Speed of light in vacuum	c	$2.9979 \times 10^{8}\,\mathrm{m/s}$
Planck constant	h	$6.626 \times 10^{-34}\,\mathrm{Js}$
Wave drag in vacuum	Γ	$376.731\,\Omega$
Stefan–Boltzmann radiation constant	σ	$5.6703 \times 10^{-8}\,\mathrm{W/(m^2\,K^4)}$
Planck radiation constants	c_1	$3.741 \times 10^{-16}\,\mathrm{W\,m^2}$
	c_2	$1.438 \times 10^{-2}\,\mathrm{m\,K}$
Wien constant	K	$2.8978 \times 10^{-3}\,\mathrm{m\,K}$
Rydberg constant	R	$1.09737 \times 10^{7}\,\mathrm{m^{-1}}$
Static weight of electrons	m_e	$9.109 \times 10^{-31}\,\mathrm{kg}$
Radius of an electron	r_e	$2.8178 \times 10^{-15}\,\mathrm{m}$
Atomic mass unit	amu	$1.6606 \times 10^{-27}\,\mathrm{kg}$

Table 18.17a,b Basic definitions (a) and constants (b) of nuclear physics

(a) Definitions

Nomenclature	Definition	Unit	Formula	Remarks
Atomic mass	The relative mass of nucleus of C_{12} is taken as unit weight	amu $= 1.6603\times10^{-27}$ kg	amu $= m_{C_{12}}/M_{C_{12}}$ $= 1/N_A$	–
Atomic number	–	–	$N = \frac{m}{N} N_A$	$N = \frac{10^{-3}\,\text{kg}}{226\,\text{kg/kmol}} 6.0221\times10^{26}\,1/\text{kmol}$ $= 2.665\times10^{21}$
Half-life period	Time required for decay of half of the original atomic mass	s, min, d, y	$T_{1/2} = \ln 2/\lambda$	${}^{238}_{92}\text{U}: T_{1/2} = 4.5\times10^{9}$ a, γ and α radiation ${}^{3}_{1}\text{H}: T_{1/2} = 2.3$ a, β radiation
Atomic energy	The energy required to move an electron through a voltage of 1 V	Electron volt $1\,\text{eV} = 1.6022\times10^{-19}$ J	$W = eU$	Nuclear fission of uranium
Mass of electron	Based on Einstein's theory of energy and mass	$1\,\text{MeV} \approx 1.728\times10^{-33}$ g	$m = \frac{E}{c_0^2}$; $m = \frac{m_0}{\sqrt{1-(c/c_0)^2}}$	$m \cong \frac{E}{c_0^2} = \frac{1.6022\times10^{-19}\,\text{J}}{(2.998\times10^{8}\,\text{m/s})^2}$ $= 1.782\times10^{-33}$ g
Absorbed energy	Amount of absorbed radiation energy per unit mass of a material	Gray $1\,\text{Gy} = 1\,\text{J/kg}$	$D = W/m$	–
Equivalent absorption	Measurement of biological radiation effect that is emitted from a γ ray of 10^{-2} Sv and is absorbed in the human body	Sievert $1\,\text{Sv} = 1\,\text{J}$	$H = DQ_F$	X-ray, β, ${}^{0}_{-1}\text{e}$, ${}^{0}_{+1}\text{e}$ radiation Quality factor Q_F Thermal neutrons 3 Alpha radiation 10
Activity	Measurement of intensity of a radioactive ray. Count of decay per unit of time	Bequerel $1\,\text{Bq} = 1/\text{s}$	–	–
Atomic cross section	Measurement of yield in a nuclear reaction. Virtual cross section of radiating atom	m^2	σ	Fission σ_f Absorption σ_a Scattering σ_s

Explanation of the table entries

${}^{A}_{Z}N : N =$ nuclear weight , $Z =$ neutron count , $A =$ atomic mass , $N = A - Z$

Fission of uranium:

$${}^{235}_{92}\text{U} + {}^{1}_{0}\text{n} \rightarrow {}^{89}_{36}\text{Kr} + {}^{144}_{56}\text{Ba} + 3\,{}^{1}_{0}\text{n} + 200\,\text{MeV}\,.$$

Energy emitted by fission of 1 g of uranium:

$$Q = \frac{m}{M} N_A W = \frac{1\,\text{g}\,6.0221\times10^{23}\,(1/\text{mol})\,200\,\text{MeV}\,1.6022\times10^{-13}\,\text{Ws/MeV}}{235\,\text{g/mol}\,3600\,\text{s/h}} = 22\,810\,\text{kW h}\,.$$

Isotopes are different types of nucleus of the same chemical elements. There nuclei have the same proton count but different atomic weights. For example:

${}^{12}_{6}\text{C}$, ${}^{13}_{6}\text{C}$, ${}^{14}_{6}\text{C}$ and ${}^{234}_{92}\text{U}$, ${}^{235}_{92}\text{U}$, ${}^{238}_{92}\text{U}$.

Types of radiation:

α : ${}^{4}_{2}\alpha$ nucleus of helium atom

β : Electrons or positrons

γ : Short-wavelength, energy-rich penetrative electromagnetic waves, which changes either the nuclear charge or the atomic mass of the radiating nucleus

Neutrons ${}^{1}_{0}\text{n}$, positrons ${}^{0}_{+1}\text{e}$, electrons ${}^{0}_{-1}\text{e}$

Table 18.17 (continued)

(b) Constants	
Velocity of light	$c_0 = 2.998 \times 10^8$ m/s
Static weight of electron	$m_{e0} = 9.110 \times 10^{-31}$ kg
Avogadro constant	$N_A = 6.0221 \times 10^{26}$ kmol^{-1}
Static weight of proton	$m_{p0} = 1.6606 \times 10^{-27}$ kg
Electric charge of an electron	$e = 1.6022 \times 10^{-19}$ C
Static weight of neutron	$m_{n0} = 1.675 \times 10^{-27}$ kg

Table 18.23 Important standards and their abbreviations

AGMA	American Gear Manufacturers Association
ANSI	American National Standard Institution
ASTM	American Society for Testing and Materials
API	American Petroleum Institute
BSI	British Standard Institution
CEN	Comité Européen de Normalisation
CENELEC	Comité Européen de Normalisation Electrotechniques
GOST	Government Standard of the former USSR
IEC	International Electrotechnical Commission
ISO	International Organization for Standardization
NF	Normes Françaises
NEN	Netherland Norms
ÖNORM	Austrian Norms
SAE	Society of Automotive Engineers
SNV	Swedish Norms
UNI	Unificazione Nazionale Italiana
DIN	Deutsches Institut für Normung

Table 18.22 Conversion from dB to pressure or power ratios and vice versa

dB	p/p_0	p^2/p_0^2
0	1	1
0.1	1.012	1.023
0.2	1.023	1.047
0.3	1.035	1.072
0.4	1.047	1.096
0.5	1.059	1.122
0.6	1.072	1.148
0.7	1.084	1.175
0.8	1.096	1.202
0.9	1.109	1.23
1.0	1.122	1.259
0	1	1
1	1.122	1.259
2	1.259	1.585
3	1.413	1.995
4	1.585	2.512
5	1.778	3.162
6	1.995	3.981
7	2.239	5.012
8	2.512	6.31
9	2.818	7.943
10	3.162	0.01
0	1	1
10	3,162	10
20	10	10^2
30	31.62	10^3
40	100	10^4
50	316.2	10^5
60	1000	10^6
70	3162	10^7
80	10000	10^8
90	31620	10^9
100	100000	10^{10}

Table 18.18 Basic units for light calculations

Measure	Definition	Unit	Mathematical relation	Remarks
Luminous flux	Amount of rays emanating from a light source in all directions	Lumen (lm)	$\phi = \mathrm{d}Q/\mathrm{d}t$	Light energy emitted per unit time
Light intensity	Intensity of light rays inside the elementary space angle[a]. 1 cd is the radiation emitted by a black body perpendicular to its surface $[1/(6\times10^6)\,\mathrm{m}^2]$ at 2042.5 K and 1.0133 bar	Candela (cd) cd = m/sr SI basic units	$I = \mathrm{d}\phi/\mathrm{d}\omega$	Stearin candle ≈ 1 cd Bulb (40 W) ≈ 35 cd
Luminance	Ratio of light emitted at the source to the light received at a particular surface	Lux (lx) $\mathrm{lx} = \mathrm{lm/m}^2$	$E = \phi/A = I\omega/A = I/r^2$	Summer sunlight 10^5 lx Living room 10–150 lx Full-moon night 0.2 lx No-moon night 3×10^{-4} lx
Light density	Light intensity per unit of illuminated surface	$\mathrm{cd/m}^2$	–	Full moon 2500 $\mathrm{cd/m}^2$ Candle 7500 $\mathrm{cd/m}^2$ Bulb 2×10^7 $\mathrm{cd/m}^2$ Sun 2.2×10^9 $\mathrm{cd/m}^2$
Light yield	Luminous flux per unit of electric power	lm/W	$\eta = \phi/P$	Tube light 44 lm/W Bulb (1000 W) 19 lm/W Bulb (40 W) 11 lm/W
Light range	Product of luminous flux and the duration of radiation	lm s	$Q = \int \phi\,\mathrm{d}t$	

[a] The unit steradian (sr) is valid for the space angles. Steradian is the ratio of the surface of a section of a sphere to the square of its radius. If α is the opening angle of a section of a sphere with an area of $A = 2\pi rh$, its height is given by $h = r[1-\cos(\alpha/2)] = 2r\sin^2(\alpha/4)$. The space angle is defined as $\omega = A/r^2 = 4\pi\sin^2(\alpha/4)$.
Special cases: $\omega = 1$ sr for $\alpha = 4\arcsin(0.5/\sqrt{\pi}) = 65.54°$. For a sphere $\alpha = 360°$ and $\omega = 4\pi$ sr. For $\alpha = 120°$ is $\omega = \pi$ sr

Table 18.20 Approximated acoustic measures

Noise source	$\eta = P_{\mathrm{acu}}/P_{\mathrm{mech}}$[a]	Noise source	$\eta = P_{\mathrm{acu}}/P_{\mathrm{mech}}$
Siren		Diesel engine	
with funnel	$3–7\times10^{-1}$	Cylinder at 800 rpm	4.00×10^{-7}
without funnel	1.00×10^{-2}	Cylinder at 3000 rpm	5.00×10^{-6}
Rotating disk with ultrasonic velocity	2.50×10^{-1}	Exhaust with turbocharger	1.00×10^{-4}
Schmidt tube	2.00×10^{-2}	Electromagnetic loudspeaker	5.00×10^{-2}
Ventilator optimal point		Electric motor	
$\Delta p < 2.5$ mbar[a]	1.00×10^{-6}	Special low noise	2.00×10^{-8}
$\Delta p > 2.5$ mbar	$4\times10^{-8}\Delta p$	Normal	1.00×10^{-6}
Escape noise		Machines	
Ma < 0.3[a]	$8(1\times10^{-6}–1\times10^{-5})\,(\mathrm{Ma})^3$	Special class	3.00×10^{-8}
0.4 < Ma < 1.0	$1.0\times10^{-4}\,(\mathrm{Ma})^5$	Quiet machines	2.00×10^{-7}
Ma > 2.0	2.00×10^{-3}	Normal	2.00×10^{-7}
		Bad	3.00×10^{-6}
Motor bike 250 cm^3 capacity		Airplane propeller	
without damper	1.00×10^{-3}	2700 kW in test stand	5.00×10^{-3}
Organ	$1\times10^{-3}–1\times10^{-2}$	Human voice	5.00×10^{-4}
Small gas turbine		Ship propeller	
Suction	1.00×10^{-4}	Without cavitation	$1\times10^{-9}–1\times10^{-8}$
Exhaust	1.00×10^{-5}	With cavitation	1.00×10^{-7}
Housing	1.00×10^{-6}		

[a] Δp = compression, P_{acu} = acoustic power, Ma = mach number, P_{mech} = mechanical power

Table 18.19 Important terms in acoustic technology

Term	Definition	Mathematical formula	Unit	Remarks
Velocity of sound	Solids			
	Longitudinal waves in big bodies	$c_L = \sqrt{\frac{2G(1-\upsilon)}{\varsigma(1-2\upsilon)}}$	m/s	1000–5000 m/s
	Transversal waves in big bodies	$c_T = \sqrt{\frac{G}{\varsigma}}$		500–3500 m/s
	Bending waves in bars	$c_D = \sqrt{\frac{E}{\varsigma}}$		Rubber 50 m/s
	Liquids	$c = \sqrt{\frac{\chi}{\varsigma}}$		Water 1485 m/s
	Gases	$c = \sqrt{xRT}$		Air: 331 m/s at 1 bar, 0 °C Hydrogen: 1280 m/s at 1 bar, 0 °C
Transverse vibrational speed of sound	Speed of vibrational part	$u = a_0\omega = 2\pi a_0 f$	m/s	5×10^{-8}–1 m/s
Sound pressure	Static and dynamic pressure in elastic media	p	N/m^2	10^{-2}–10^2 N/m^2 Normal audio waves $= 2\times10^{-5}$ N/m^2 Piano $= 0.2$ N/m^2 Siren $= 35$ N/m^2
Sound power	Sound energy that passes per unit time through a particular surface area	P	W	1×10^{-12}–1×10^5 W Audio waves $= 1\times10^{-12}$ W Voice $\approx 1\times10^{-3}$ W Siren $\approx 1\times10^3$ W
Sound intensity	Sound power per unit area	$I = P/A = p^2/cp$	W/m^2	1×10^{-11}–1×10^3 W/m^2 Audio waves $= 1\times10^{-12}$ W/m^2
Sound level	Logarithmic scale for sound pressure	$L = 10\lg(P/P_0)$		0–140 dB
		$= 10\lg(I/I_0)$	Bel	$P_0 = 1\times10^{-12}$ W $I_0 = 1\times10^{-12}$ W/m^2
		$= 20\lg(p/p_0)$	B, dB	$P_0 = 2\times10^{-5}$ N/m^2
Sound volume	Measurement of subjective perception of the sound intensity for the ear	$\Lambda = 10\lg(I/I_0)$	phon	0–130 phon Audio wave 0 phon Entertainment 50 phon Pain level 130 phon
Sound absorption factor	Measurement of loss of sound energy in heat due to friction	$\alpha = (P_a - P_f)/P_f$	1	For 500 Hz
	Index a and r indicate absorption and reflection, respectively	$= (p_a^2 - p_r^2)/p_r^2$		Concrete 0.01 Glass 0.03 Foam 0.36
Sound damping	Logarithmic measurement for sound damping by a wall, Index 1 indicates energy before reflection and 2 indicates energy after reflection	$R = 10\lg(I_1/I_2)$	dB	1 mm-thick steel plate 29 dB
Acoustic coefficient	Ratio of sound to mechanical power	$\eta = P_{acu}/P_{mech}$	1	See Table 18.20

a_0 Amplitude, A surface area, P power, x isentropic exponent, f frequency, E modulus of elasticity, R gas constant, υ Poisson ratio, G modulus of rigidity, T absolute temperature, ς density, χ compressibility

Table 18.21 Periodic table of chemical elements

Group	1	2		3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Period																			
1	1 H Hydrogen																		2 He Helium
2	3 Li Lithium	4 Be Beryllium												5 B Boron	6 C Carbon	7 N Nitrogen	8 O Oxygen	9 F Fluorine	10 Ne Neon
3	11 Na Sodium	12 Mg Magnesium												13 Al Aluminium	14 Si Silicon	15 P Phosphorus	16 S Sulphur	17 Cl Chlorine	18 Ar Argon
4	19 K Potassium	20 Ca Calcium		21 Sc Scandium	22 Ti Titanium	23 V Vanadium	24 Cr Chromium	25 Mn Manganese	26 Fe Iron	27 Co Cobalt	28 Ni Nickel	29 Cu Copper	30 Zn Zinc	31 Ga Gallium	32 Ge Germanium	33 As Arsenic	34 Se Selenium	35 Br Bromine	36 Kr Krypton
5	37 Rb Rubidium	38 Sr Strontium		39 Y Yttrium	40 Zr Zirconium	41 Nb Niobium	42 Mo Molybdenum	43 Tc Technetium	44 Ru Ruthenium	45 Rh Rhodium	46 Pd Palladium	47 Ag Silver	48 Cd Cadmium	49 In Indium	50 Sn Tin	51 Sb Antimony	52 Te Tellurium	53 I Iodine	54 Xe Xenon
6	55 Cs Caesium	56 Ba Barium	*	71 Lu Lutetium	72 Hf Hafnium	73 Ta Tantalum	74 W Tungsten	75 Re Rhenium	76 Os Osmium	77 Ir Iridium	78 Pt Platinum	79 Au Gold	80 Hg Mercury	81 Tl Thallium	82 Pb Lead	83 Bi Bismuth	84 Po Polonium	85 At Astatine	86 Rn Radon
7	87 Fr Francium	88 Ra Radium	**	103 Lr Lawrencium	104 Rf Rutherfordium	105 Db Dubnium	106 Sg Seaborgium	107 Bh Bohrium	108 Hs Hassium	109 Mt Meitnerium	110 Ds Darmstadtium	111 Rg Roentgenium	112 Uub Ununbium	113 Uut Ununtrium	114 Uuq Ununquadium	115 Uup Ununpentium	116 Uuh Ununhexium	117 Uus Ununseptium	118 Uuo Ununoctium
*Lanthanoids			*	57 La Lanthanum	58 Ce Cerium	59 Pr Praseodymium	60 Nd Neodymium	61 Pm Promethium	62 Sm Samarium	63 Eu Europium	64 Gd Gadolinium	65 Tb Terbium	66 Dy Dysprosium	67 Ho Holmium	68 Er Erbium	69 Tm Thulium	70 Yb Ytterbium		
**Actinoids			**	89 Ac Actinium	90 Th Thorium	91 Pa Protactinium	92 U Uranium	93 Np Neptunium	94 Pu Plutonium	95 Am Americium	96 Cm Curium	97 Bk Berkelium	98 Cf Californium	99 Es Einsteinium	100 Fm Fermium	101 Md Mendelevium	102 No Nobelium	

Acknowledgements

B.6 Design of Machine Elements

by Oleg P. Lelikov

The author would like to thank Dr. Juri Postnikov for his great collaboration and efforts in the preparation of this chapter.

B.7 Manufacturing Engineering

by Thomas Böllinghaus, Gerry Byrne, Boris Ilich Cherpakov (deceased), Edward Chlebus, Carl E. Cross, Berend Denkena, Ulrich Dilthey, Takeshi Hatsuzawa, Klaus Herfurth, Horst Herold (deceased), Andrew Kaldos, Thomas Kannengiesser, Michail Karpenko, Bernhard Karpuschewski, Manuel Marya, Surendar K. Marya, Klaus-Jürgen Matthes, Klaus Middeldorf, Joao Fernando G. Oliveira, Jörg Pieschel, Didier M. Priem, Frank Riedel, Markus Schleser, A. Erman Tekkaya, Marcel Todtermuschke, Anatole Vereschaka, Detlef von Hofe, Nikolaus Wagner, Johannes Wodara, Klaus Woeste

The author would like to express his thanks to all his colleagues from leading welding institutions and departments for their deep interest and active cooperation in revising the section “Joining Technology”, and to representatives of the Deutscher Verband für Schweißen und verwandte Verfahren e.V. (German Association for Welding and Related Processes) in Düsseldorf for writing the essential analytical introduction.

In particular, I owe thanks to my colleague Dr.-Ing. J. Pieschel from the Institute of Joining and Beam Technology at the Otto von Guericke University in Magdeburg for his careful editorial preparation of the manuscript.

About the Authors

Gritt Ahrens

Daimler AG X944
Systems Integration and Comfort Electric
Sindelfingen, Germany
gritt.ahrens@daimler.com

Chapter B.13, Sect. 13.2

Gritt Ahrens received the PhD degree from the Department of Engineering Design and Methodology at the Technical University of Berlin. She developed a method for capture and management of product requirements and substantiated its practical applicability. Since 2001 she has been developing and optimizing various processes in the context of car development at Mercedes-Benz Cars Development (Daimler AG).

Seddik Bacha

Université Joseph Fourier
Grénoble Electrical Engineering
Laboratory
Grenoble, France
seddik.bacha@g2elab.inpg.fr

Chapter C.17, Sect. 17.4

Professor Seddik Bacha received the Engineering and Magister degree from École Nationale Polytechnique of Algiers in 1982 and 1990, and the PhD and HDR degrees in 1993 and 1998 from Polytechnic Institute of Grenoble, France, respectively. He is presently the manager of the Power System Group in the Grénoble Electrical Engineering Laboratory and Professor at the Université Joseph Fourier, Grenoble. His interests are power electronics systems, power quality, and renewable energy.

Stanley Baksi

TRW Automotive, Lucas Varity GmbH
Koblenz, Germany
stanley.baksi@trw.com

Chapter C.18

Stanley Baksi received the BA degree in Engineering in 2000 from Visvesvaraya National Institute of Technology and the MSc in Mechanical Engineering from California State University, Long Beach. He completed his doctoral research on rapid bone reconstruction using reverse engineering at Otto-von-Guericke University, Magdeburg, in 2007. Since January 2008 he has been a Project Engineer with TRW Automotive providing computer-aided engineering support for the development of automotive brakes.

Thomas Böllinghaus

Federal Institute for Materials Research and Testing (BAM)
Berlin, Germany
thomas.boellinghaus@bam.de

Chapter B.7, Sect. 7.4.8

Professor Thomas Böllinghaus received the Dr.-Ing. degree in 1995 from the Helmut-Schmidt-University, Hamburg, and qualified there as a Professor for Materials Science and Welding Technology in 1999. He then became head of Division V.5 Safety of Joined Components at the Federal Institute for Materials Research and Testing (BAM), Berlin. Since 2003 he has been Vice President of BAM and became Honorary Professor for Failure Analysis and Failure Prevention in the Faculty for Mechanical Engineering at the Otto-von-Guericke University, Magdeburg in 2008. His research interests focus on hydrogen-assisted cracking of structural metallic materials, numerical simulation of hydrogen transport and cracking, hot cracking phenomena in welds, and failure analysis and lifecycle evaluations of joined components.

Alois Breiing

Eidgenössische Technische Hochschule
Zürich (ETH)
Institut für mechanische Systeme (IMES)
Zentrum für Produkt-Entwicklung (ZPE)
Zurich, Switzerland
breiing@imes.mavt.ethz.ch

Chapter B.9, Sect. 9.6

Alois Breiing is a retired Professor (Titularprofessor) at ETH Zürich. He was a Lecturer (Privatdozent) at the Institute of Mechanical Systems. He has 23 years of experience as an aerospace engineer in the Development and Design Department of the aerospace company Dornier. Beginning in 1985 he worked at the Institute of Design and Construction Methods of ETH Zurich (now the Institute of Mechanical Systems), starting as a Senior Researcher and Section Leader. He received a doctor's degree in 1991. He gave lectures in machine design. The main topics of his research work were evaluation and decision processes.

Eugeniusz Budny

Institute of Mechanized Construction and Rock Mining
Warsaw, Poland
e.budny@imbigs.org.pl

Chapter B.14, Sect. 14.1

Dr. Eugeniusz Budny is a Professor of Mechanical Engineering and Emeritus Director of the Institute of Mechanized Construction and Rock Mining (IMBiGS) in Warsaw, Poland. He began his career as an automotive industry product designer and then moved to the construction equipment manufacturing industry, specializing in the design of single-bucket excavators, hydraulic drives, and hydraulic engineering machine controls. Since 1977 he has been engaged in the mechanization of construction work. Professor Budny served as President of the International Association for Automation and Robotics in Construction (IAARC) and has been Chair of ISO/TC 195 and the Polish Committee for Standardization in the field of building construction machinery since 2005.

Gerry Byrne

University College Dublin
School of Electrical, Electronic and Mechanical Engineering
Belfield, Dublin 4, Ireland
gerald.byrne@ucd.ie

Chapter B.7, Sects. 7.3.1, 7.3.2

Gerry Byrne F.R.Eng. is the Professor of Mechanical Engineering and Director of the Advanced Manufacturing Science Research Centre at University College Dublin (UCD), Ireland. He has over 30 years of experience with high-precision manufacturing processes, surface engineering, and aspects of surface microstructuring. He is an expert in the field of cutting processes and has published extensively. He is a Fellow of the International Academy for Production Engineering (CIRP) and a Member of the Council of that Academy. He is an International Fellow of the Royal Academy of Engineering, UK, and of the Society of Manufacturing Engineers, USA.

Edward Chlebus

Wrocław University of Technology
Centre for Advanced Manufacturing Technologies
Wrocław, Poland
edward.chlebus@pwr.wroc.pl

Chapter B.7, Sect. 7.5

Professor Edward Chlebus is Director of the Institute of Production Engineering and Automation at Wrocław University of Technology, Poland, and the Head of the Centre for Advanced Manufacturing Technologies (CAMT). His main research areas are design methodology and CA*x* and PDM/PLM systems, rapid prototyping, reverse engineering, modeling, optimization, and simulation of production processes. He is a Contractor for six International Projects in FP6, Leonardo da Vinci, and ERA Net. He is also a Graduate of the University of Connecticut Business School.

Mirosław Chłosta

IMBiGS – Institute for Mechanized Construction and Rock Mining (IMBiGS)
Warsaw, Poland
m.chlosta@imbigs.org.pl

Chapter B.14, Sect. 14.1

Miroslaw Chlosta received the M.S. degree from Warsaw Technical University in 1980. In that year he began work at IMBiGS as a Research Associate. He received the PhD degree from IMBiGS in 2003 and was an Assistant there from 2003 onwards. His research interests are mainly in CAD and robotics in the building industry.

Norge I. Coello Machado

Universidad Central "Marta Abreu" de Las Villas
Faculty of Mechanical Engineering
Santa Clara, Cuba
norgec@uclv.edu.cu

Chapter B.8, Sect. 8.1

Norge Coello Machado received the Dipl.-Ing. degree from the Universidad of Santa Clara in 1982 and the Dr.-Ing. degree from the University of Magdeburg, Germany, in 1989. From 2003 he was Temporary Professor at the University of Magdeburg, Germany. His research interests are mainly in the use of statistical methods in quality management, quality engineering, and measurement technology.

Francesco Costanzo

Alenia Aeronautica
Procurement/Sourcing Management
Department
Pomigliano (NA), Italy
fcostanzo@inwind.it

Chapter B.15, Sect. 15.8

Graduated in Aeronautical Engineering in 1977 in Naples, Italy, Francesco Costanzo has a long experience in design and industrialization processes at Alenia Aeronautica (formerly Aeritalia) until 1983. He was then in charge of process and information management with responsibility for CAE/CAD/CAM until 2000. Later he was responsible for the information systems of the A380 program in Alenia and has been head of Procurement Direction/Sourcing Management as Supplier Value Creation Manager, with focus on processes and information for Alenia's supply chain. Currently Costanzo is a Senior Consultant in processes for procurement and the digitization of supply chains.

Carl E. Cross

Federal Institute for Materials Research
and Testing (BAM)
Joining Technology
Berlin, Germany
carl-edward.cross@bam.de

Chapter B.7, Sect. 7.4.8

Carl Cross received the B.Sc., MSc, and PhD degrees in Metallurgical Engineering from the Colorado School of Mines. He has worked in welding research for over 30 years in both industry and academia, specializing in solidification defects and weldability testing of nonferrous alloys. Currently he is a Senior Scientist at BAM (the German Federal Institute for Materials Research and Testing), investigating the critical conditions needed to initiate and grow hot cracks.

Frank Dammel

Technical University
Department of Mechanical Engineering/
Institute of Technical Thermodynamics
Darmstadt, Germany
dammel@ttd.tu-darmstadt.de

Chapter B.4, Sect. 4.3

Frank Dammel is a postdoctoral scientific assistant (Akad. Oberrat) at the Institute of Technical Thermodynamics at the Technische Universitaet Darmstadt. His main field of research is phase-change phenomena with an emphasis on numerical simulations.

Jaime De La Ree

Virginia Tech
Electrical and Computer Engineering
Department
Blacksburg, VA, USA
jreelope@vt.edu

Chapter C.17, Sect. 17.3

Jaime De La Ree received the PhD degree from the University of Pittsburgh, Pittsburgh, PA, USA in 1984. He worked at the ECE Department at Virginia Tech and became Assistant Department Head in 2004. His teaching and research interest is in the areas of electric machinery and power systems protection. He is the recipient of several Outstanding Teaching Awards from both the University of Pittsburgh and Virginia Tech.

Torsten Dellmann

RWTH Aachen University
Department of Rail Vehicles and
Materials-Handling Technology
Aachen, Germany
torsten.dellmann@ifs.rwth-aachen.de

Chapter B.13, Sect. 13.1

Dr. Torsten Dellmann is a Professor of Rail Vehicles and Materials-Handling Technology at RWTH Aachen University, Germany. He obtained the Dr. degree in materials-handling technology from RWTH Aachen in 1986. Prior to joining the RWTH Aachen University as a professor he worked as Managing Director for Knorr-Bremse MRP Systems for Rail Vehicles. His current research is focused on innovative braking systems as well as innovative undercarriages and guiding technology for rail vehicles.

Berend Denkena

Leibniz University Hannover
IFW – Institute of Production Engineering and Machine Tools
Garbsen, Germany
denkena@ifw.uni-hannover.de

Chapter B.7, Sects. 7.3.1, 7.3.2, 7.3.3

Professor Denkena earned the PhD degree at Leibniz University Hanover, Germany, in 1992. He then became Manager Standards Engineering at Thyssen Production Systems in Auburn Hills, USA (1993–1995). Back in Germany he served as Manager Mechanical Development at Thyssen Hüller Hille in Ludwigsburg for 1 year before he went to Bielefeld as Manager Engineering and Development for Gildemeister Turning Machines. Since 2001 he has been a Professor at the Leibniz University Hannover and the Head of the Institute of Production Engineering and Machine Tools.

Ludger Deters

Otto-von-Guericke University
Institute of Machine Design
Magdeburg, Germany
ludger.deters@ovgu.de

Chapter B.5

Ludger Deters is Professor of Machine Elements and Tribology at the Otto-von-Guericke University of Magdeburg. He received the PhD degree from the Mechanical Engineering Department of the University of Clausthal in 1983. From 1983 to 1994 he worked in executive positions in different development and design departments of the machine-building industry. His current main research activities include tribology, slider and rolling bearings, wheel/rail contacts, and friction and wear of combustion engine parts.

Ulrich Dilthey

RWTH Aachen University
ISF Welding and Joining Institute
Aachen, Germany
di@isf.rwth-aachen.de

Chapter B.7, Sects. 7.4.3, 7.4.4, 7.4.7

Professor Dr.-Ing. Ulrich Dilthey holds degrees in Mechanical Engineering/Production Engineering and a doctorate from RWTH Aachen University, where he was research assistant in the ISF, the Welding and Joining Institute of the RWTH Aachen. He has many years of working experience in leading positions of worldwide operating companies in the fields of mechanization, automation, and robotics. From 1989 to 2007 he was Professor of Welding Manufacturing Processes and Director of the ISF. In July 2008 he was elected the President of the International Institute of Welding (IIW).

Frank Engelmann

University of Applied Sciences Jena
Department of Industrial Engineering
Jena, Germany
frank.engelmann@fh-jena.de

Chapter B.9, Sects. 9.1–9.4

Frank Engelmann studied mechanical engineering at the Engineering Department at the University of Magdeburg. After he finished his studies he worked as Managing Director for a production business while also working at the University of Magdeburg, where he received the PhD degree. His research activities focus on secondary explosion protection and on biomedical technology. In October 2007 Frank Engelmann joined the University of Applied Sciences in Jena, Germany, as a full professor.

Ramin S. Esfandiari

California State University
Department of Mechanical & Aerospace Engineering
Long Beach, CA, USA
esfandi@csulb.edu

Chapter A.1

Dr. Esfandiari is Professor of Mechanical and Aerospace Engineering at the California State University, Long Beach. He received the PhD degree from the University of California at Santa Barbara in Applied Mathematics with emphasis on optimal control theory. He has published numerous research papers in refereed journals in the areas of distributed control and numerical methods. He is the author/coauthor of three textbooks and author of a MATLAB manual.

Jens Freudenberger

Leibniz-Institute for Solid State and Materials Research Dresden
Department for Metal Physics
Dresden, Germany
j.freudenberger@ifw-dresden.de

Chapter B.3, Sect. 3.1

Dr. Jens Freudenberger is Head of the Department for Metal Physics and Supervisor of the Pulsed High-Magnetic-Field Laboratory at the IFW, Dresden. His current research is dedicated to plastic deformation and metal forming. He is an expert in materials science and solid-state physics with particular interest in materials such as superconductors, high-strength nanostructured conductors, and magnetic materials.

Stefan Gies

RWTH Aachen University
Institute for Automotive Engineering
Aachen, Germany
gies@ika.rwth-aachen.de

Chapter B.13, Sect. 13.1

Professor Dr.-Ing. Stefan Gies received the doctor degree in 1993 from Aachen University at the Institute for Automotive Engineering (IKA), where he was employed as a Scientific Research Engineer between 1989 and 1993 and Chief Engineer from 1993 until 1994. From 1994 to 2000, he was with Ford AG and changed in 2000 to Audi AG, where he was Manager of vehicle dynamics and driving comfort with steering systems, suspension, wheels/tires, chassis controls. Since 2007, he has been a Professor at RWTH Aachen University and the head of IKA.

Joachim Göllner

Otto-von-Guericke University
Institute of Materials and Joining Technology Department of Mechanical Engineering
Magdeburg, Germany
joachim.goellner@mb.uni-magdeburg.de

Chapter B.3, Sect. 3.6

Dr. Joachim Göllner is a Lecturer for materials and corrosion. He habilitated from Otto-von-Guericke University Magdeburg, Germany, and is the Head of a research group. His research interests are in electrochemical noise and the development of corrosion test methods. He has authored more than 150 publications and is a member of various professional societies.

Timothy Gutowski

Massachusetts Institute of Technology
Department of Mechanical Engineering
Cambridge, MA, USA
gutowski@mit.edu

Chapter B.9, Sect. 9.5

Timothy Gutowski is a Professor of Mechanical Engineering at the Massachusetts Institute of Technology. He received the PhD degree in Mechanical Engineering from MIT, the M.S. degree in Theoretical and Applied Mechanics from the University of Illinois, and the B.Sc. degree in Mathematics from the University of Wisconsin. His current research interests focus on the relationship between manufacturing and the environment. He was Director of MIT's Laboratory for Manufacturing and Productivity for 10 years and Associate Department Head of Mechanical Engineering from 2001 to 2005.

Takeshi Hatsuzawa

Tokyo Institute of Technology
Precision and Intelligence Laboratory
Yokohama, Japan
hat@pi.titech.ac.jp

Chapter B.7, Sect. 7.6

Takeshi Hatsuzawa received the Dr. Eng. degree from Tokyo Institute of Technology in 1993. After 12 years at the National Research Laboratory of Metrology, he became Associate Professor in the Tokyo Institute of Technology in 1995, and full professor in 2002. His current research is focused on MEMS and NEMS applications in biochemistry using biotechnology.

Markus Hecht

Berlin University of Technology
Institute of Land and Sea Transport Systems Department of Rail Vehicles
Berlin, Germany
markus.hecht@tu-Berlin.de

Chapter B.13, Sect. 13.3

Marskus Hecht obtained his industrial experience at the Swiss Locomotive and Machine Works Winterthur, working on radially steered bogies and low-noise equipment. Since 1997 he has been Professor for Rail Vehicles at Berlin University of Technology (TU). His research fields are wheel–rail interaction, vehicle design, vehicle acoustics, and safety. He works in several EU projects and is a member of various advisory boards. Since 2005 he has been Director of the Institute of Land and Sea Transportation at TU Berlin. He is responsible for the vehicle-related part of the postgraduate master study on Public Transport Management at Ruhr Campus, Essen, and is Advisory Professor at Tongji University Shanghai.

Hamid Hefazi

California State University
Mechanical and Aerospace Engineering
Department of Mechanical and
Aerospace Engineering
Long Beach, CA, USA
hefazi@csulb.edu

Chapter B.13, Sect. 13.4

Hamid Hefazi received the PhD degree in Aerospace Engineering from the University of Southern California in 1985. He is currently Professor and Chair of mechanical and aerospace engineering and Director of the Boeing Technology Center at California State University, Long Beach. His research activities include computational fluid dynamics (CFD), aerodynamic design optimization, aeroacoustics, hydrodynamics, neural networks, and advanced optimization methods.

Martin Heilmaier

Technical University
Department of Physical Metallurgy
Darmstadt, Germany
m.heilmaier@phm.tu-darmstadt.de

Chapter B.3, Sects. 3.1, 3.4

Professor Martin Heilmaier graduated in Materials Science from the University of Erlangen-Nuremberg. From 2002 he was Professor for Materials Testing at the Otto-von-Guericke University, Magdeburg. Since October 2008 he holds the chair for Physical Metallurgy at the TU Darmstadt. His research is dedicated to the synthesis and properties of structural materials such as refractory metal silicide alloys and intermetallic alloys for high-temperature applications.

Rolf Henke

RWTH Aachen University
Institute of Aeronautics and Astronautics
Aachen, Germany
henke@ilr.rwth-aachen.de

Chapter B.13, Sect. 13.1

Since 2006, Professor Rolf Henke has been Head of the Institute of Aeronautics and Astronautics at RWTH Aachen University. He has 20 years of experience working for Airbus R&T, in charge of large technology projects at the aircraft level including flight testing. At Aachen University, his current research projects include seamless air transport, conceptual aircraft design, development of high-lift systems, investigation of wake vortices, and aircraft component and trajectory noise. He is a member of the executive board of the German aerospace society (DLR).

Klaus Herfurth

Industrial Advisor
Langenfeld, Germany
klaus.herfurth@t-online.de

Chapter B.7, Sect. 7.1

Professor Dr.-Ing. habil. Klaus Herfurth is an expert from the foundry industry. He earned the Dr.-Ing. degree from Technische Universitat Bergakademie Freiberg, Germany in 1963, and his Habilitation in 1979. He is author of 132 technical papers and book and handbook chapters. He was a Teacher at Technische Universitat Chemnitz from 1968 to 2002 in foundry technology and materials science. From 1987 to 2001 he was Manager of the Iron and Steel Casting Technical Group and the Standardization Technical Group of the German Foundrymen's Association in Düsseldorf. He also worked in R&D at this time in the field of analysis of energy demand in foundries and realization of material and energy savings through castings.

Chris Oliver Heyde

Otto-von-Guericke University
Electric Power Networks and Renewable
Energy Sources
Magdeburg, Germany
chris.heyde@ovgu.de

Chapter C.17, Sect. 17.2

Chris O. Heyde studied electrical engineering at the Otto-von-Guericke University Magdeburg, Germany. He graduated in 2005 with the Dipl. -Ing. degree. He then joined the Chair of Electric Power Networks and Renewable Energy Sources at the Otto-von-Guericke University Magdeburg, Germany as a Research Engineer in 2005. His primary field of interest is network stability.

Andrew Kaldos

AKM Engineering Consultants
Bebington, Wirral, UK
andrew.kaldos@ntlworld.com

Chapter B.7, Introduction

Andrew Kaldos received the Dipl.-Ing. degree in Mechanical and Production Engineering and Electrical, Instrumentation, and Control Engineering (1966, 1971) and the Dr.- Ing. degree in 1974 from the Technical University of Budapest, Hungary. He then held various positions at the Technical University of Budapest, from 1986 to 1989 at Leeds Polytechnic, from 1989 to 2000 at John Moores University, Liverpool, and since 2007 he has been Managing Director of AKM Engineering Consultants. He is involved in industrial technology transfer in the area of manufacturing engineering by research cooperation with eight departments of six universities in Europe, and development of cooperation with industrial companies.

Yuichi Kanda

Toyo University
Department of Mechanical Engineering
Advanced Manufacturing Engineering Laboratory
Kawagoe-City, Japan
cimkanda@toyonet.toyo.ac.jp

Chapter B.15, Sect. 15.6

Yuichi Kanda, is a Professor in the Faculty of Engineering at Toyo University, Japan where he received the M.S. and PhD degrees in Mechanical Engineering. His area of research is advanced manufacturing systems using networks and ultraprecision machining. He is a Fellow of the Japan Society of Mechanical Engineering and is the Chairperson of the FA Control Network committee of JEMA and the Society of Project Management.

Thomas Kannengiesser

Federal Institute for Materials Research and Testing (BAM)
Joining Technology
Berlin, Germany
thomas.kannengiesser@bam.de

Chapter B.7, Sect. 7.4.8

Thomas Kannengiesser studied Materials Engineering at Magdeburg University, Faculty of Mechanical Engineering. He was a scientific employee within the scope of a doctoral candidate program of BAM Federal Institute for Materials Research and Testing, Berlin, Germany. Since 2005, he has been Head of the Working Group Testing of Welded Components. His main fields of activity are component weld tests and full-scale tests under defined restraint intensities, cold cracking investigations on small and large specimens, and mobile residual stress measurement on components.

Michail Karpenko

New Zealand Welding Centre
Heavy Engineering Research Association (HERA)
Manukau City, New Zealand
michail.karpenko@hera.org.nz

Chapter B.7, Sect. 7.4.2

Dr. Karpenko obtained the Mechanical Engineer degree from the Kiev Polytechnic Institute in 1995. He extended this qualification with the Welding Engineer's degree (IWE). He completed the PhD degree at the Otto-von-Guericke University, Magdeburg, Germany, in 2001. From 1996 to 2006 he was a Research Fellow at the University of Magdeburg involved in a number of research projects in laser welding, drilling, and hybrid and plasma welding. In 2006 Dr. Karpenko took over the position of the NZ Welding Centre Manager in Manukau, New Zealand.

Bernhard Karpuschewski

Otto-von-Guericke University
Department of Manufacturing Engineering
Magdeburg, Germany
karpu@mb.uni-magdeburg.de

Chapter B.7, Sects. 7.3.1, 7.3.2, 7.3.3

After receiving the PhD degree and holding the Chief Engineer position at the University of Hannover, Germany, Bernhard Karpuschewski worked for 1.5 years as Associate Professor at Keio University, Japan, and then for 4.5 years as a Full Professor for Production Technology at TU Delft, The Netherlands. Since 2005 he has been Full Professor and Managing Director of IFQ in Magdeburg. He is a Fellow of the International Academy for Production Engineering (CIRP), and the current Chairman of Abrasive Processes.

Toshiaki Kimura

Japan Society for the Promotion of Machine Industry (JSPMI)
Production Engineering Department
Technical Research Institute
Tokyo, Japan
kimura@tri.jspmi.or.jp

Chapter B.15, Sect. 15.6

Toshiaki Kimura is an Engineering Leader of the Production Engineering Department at the Technical Institute of the Japan Society for the Promotion of Machine Industry (JSPMI). His interests and research areas are methodologies and application systems of manufacturing support systems using information technologies. He is a member of the JSME and the JSPE.

Dwarkadas Kothari

VIT University
School of Electrical Sciences
Vellore, TN, India
dkothari@ces.iitd.ac.in

Chapter C.16

Prof. D.P. Kothari is Former Director (in charge) of IIT, Delhi. He is currently Vice Chancellor of VIT University, Vellore. He is FNAE, FNASc, SMIEEE, MIEE, FIE (India), National Khosla Award Winner for Life Time Achievements in engineering. He has co-authored 20 books and more than 540 research publications, and has guided 28 PhD and 59 masters theses. His area of research interest is power and energy system engineering, operation, control, reliability, and optimization.

Hermann Kühnle

Otto-von-Guericke University
Institute of Ergonomics Factory
Operations and Automation
Magdeburg, Germany
hermann.kuehnle@ovgu.de

Chapter B.15, Sects. 15.1, 15.2, 15.4, 15.5, 15.9

Hermann Kühnle is Full Professor for Factory Operation and Manufacturing Systems and Executive Director at the Otto-von-Guericke University of Magdeburg. Hermann Kühnle has worked on approaches and concepts for high-performance organizations on the base of fractal geometry and complexity theory, and implemented these structures in many world market-leading companies. He has been involved in numerous research projects with international partners of global dimensions. Hermann Kühnle is a member of several international scientific committees.

Oleg P. Lelikov

Bauman Moscow State Technical University
Moscow, Russia

Chapter B.6

Professor Lelikov received the degree as Mechanical Engineer from Bauman Moscow State Technical University (BMSTU) in 1965, and in 1969 the academic status of Candidate of Technical Sciences. Since 1971 he has been working in the Department of Design Principles of Machines. Since 1992 he has been Professor at BMSTU, carrying out research on power-train loading, roadholds of multiaxial and multi-drive cars, safety of the multiple-unit reducer mechanisms, efficiency of gearings, ballscrew assemblies, and computer-aided design of machine components. He is an editor of one of the volumes of the Russian mechanical engineering encyclopedia in 40 volumes.

Andreas Lindemann

Otto-von-Guericke University
Institute for Power Electronics
Magdeburg, Germany
andreas.lindemann@ovgu.de

Chapter C.17, Sect. 17.4

Andreas Lindemann holds the Dr.-Ing. degree in Electrical Engineering. For 10 years he was responsible for the development of power semiconductor components in industry. Being appointed Professor of Power Electronics at Helmut-Schmidt-Universität, Hamburg, in 2004, he now holds the Chair for Power Electronics at Otto-von-Guericke University Magdeburg, Germany.

Bruno Lisanti

AST
Lonate Pozzolo (VA), Italy
bruno.lisanti@ast-italia.com

Chapter B.15, Sect. 15.8

Partner and Principal Consultant at AST in Italy, Bruno Lisanti has more than 25 years of experience in technical and management consulting in a wide range of markets (aerospace and defense, energy, automotive, and engineering) and company sizes, including large, small, and medium-sized organizations. Graduating in Nuclear Engineering from the University of Pisa, he is directly involved in consulting services, and in the last 10 years has participated in and coordinated several European research projects on concurrent engineering, virtual organizations, and advanced collaborative approaches, methods, and tools.

Manuel Marya

Schlumberger Reservoir Completions
Material Engineering
Rosharon, TX, USA
mmarya@slb.com

Chapter B.7, Sect. 7.4.9

Dr. Marya is Group Leader with Schlumberger Technology Corp., TX. After completing his education in France, Canada, and the USA, where he received the M.S. and PhD degrees (Colorado School of Mines), he spent 3 years at General Motors R&D (MI) and NanoCoolers (TX). Dr. Marya has over 50 publications in materials design and processing to his name and is the recipient of the 2000 AWS Graduate Fellowship Award, 2002 IIW Henry Granjon Prize, and the 2006 AWS William Spraragen Award.

Surendar K. Marya

GeM–UMR CNRS 6183, Ecole Centrale Nantes
Institut de Recherche en Génie Civil et Mécanique
Nantes, France
surendar.marya@ec-nantes.fr

Chapter B.7, Sect. 7.4.9

Surendar Marya is Professor in Materials Science and Engineering, Physical Metallurgy, and Manufacturing Processes at Centrale Nantes, France. He holds the PhD degree from Punjab University College, India, and the Dr. Es. Sc. degree from the University of Paris Orsay, France. Professor Marya is a technical advisor to several industries and is heavily involved in scientific missions worldwide. He has been a Visiting Professor in the USA, Japan, and Korea, and frequently lectures in Southwest Asia.

Ajay Mathur

Simon India Limited
Plant Engineering
New Delhi, India
avm2k@vsnl.net

Chapter B.11

Ajay Mathur is a Mechanical Engineer graduated from The Maharaja Sayaji Rao University, Baroda, India. He has over 20 years of experience in design and fabrication of pressure vessels, heat exchangers, skid-mounted plants, and fired heater modules for refinery, petrochemical, nuclear, and chemical projects in India and elsewhere.

Klaus-Jürgen Matthes

Chemnitz University of Technology
Institute for Manufacturing/Welding Technology
Chemnitz, Germany
schweisstech@mb.tu-chemnitz.de

Chapter B.7, Sect. 7.4.5

Klaus-Jürgen Matthes is a Professor of Welding Engineering and Director of the Institute for Manufacturing/Welding Technology at Chemnitz University of Technology. His research interests include welding processes, engineering and design weldments, laser welding of microstructures, and automation and sensor-based monitoring of welding processes. He held the position of Vice President for Research from 1997 to 2003 and has been President of Chemnitz University of Technology since October 2003.

Henning Jürgen Meyer

Technische Universität Berlin
Berlin Institute of Technology
Konstruktion von Maschinensystemen
Berlin, Germany
henning.meyer@tu-berlin.de

Chapter B.14, Sect. 14.2

Dr. Meyer is a Professor of Machinery System Design at the Berlin Institute of Technology. He studied Mechanical Engineering at the Technische Universität Braunschweig, obtaining the PhD degree in 1998. From 1998 to 2002 he worked in the Wirtgen Group and was responsible for electronic systems in road pavers. His research work is concentrated on self-configuring mobile networks and on driving dynamics and braking systems of commercial vehicles, tractors, and mobile working machines.

Klaus Middeldorf

DVS – German Welding Society
Düsseldorf, Germany
klaus.middeldorf@dvs-hg.de

Chapter B.7, Sect. 7.4.1

Klaus Middeldorf graduated as a Mechanical Engineer (1982) and received the Doctorate in Materials Science (1986) from the University of Essen. He worked as Project Manager for Paper Production for Procter &Gamble, then took the position of R&D Managing Director at the Federation of Industrial Research Associations, Cologne (1988–1999). He currently serves as R&D Managing Director, and since 2006 as General Manager, for the German Welding Society (DVS).

Gerhard Mook

Otto-von-Guericke University
Department of Mechanical Engineering
Institute of Materials and Joining Technology and Materials Testing
Magdeburg, Germany
mook@mb.uni-magdeburg.de

Chapter B.3, Sect. 3.5

Professor Mook's main area of research is nondestructive testing and evaluation as well as structural health monitoring. His work is focused on high-resolution imaging systems suitable for in-field application in aerospace and ground transportation systems. He is the chairman of the Board of University Teachers of the German Society for Nondestructive Testing (DGZfP).

Jay M. Ochterbeck

Clemson University
Department of Mechanical Engineering
Clemson, SC, USA
jochter@clemson.edu

Chapter B.4, Sect. 4.2

Jay M. Ochterbeck has been a Professor of Mechanical Engineering at Clemson University since 1994. His main fields of research are multiphase flows and heat transfer, capillary-driven loops, heat pipe science and technology, and thermal contact conductance. He is an Associate Fellow of the AIAA and is a member of several professional societies and international committees in thermal sciences.

Joao Fernando G. Oliveira

University of São Paulo
Department of Production Engineering
São Carlos, SP, Brazil
jfgo@sc.usp.br, presidencia@ipt.br

Chapter B.7, Sect. 7.3.3

Dr. Oliveira is Professor of Production Engineering at the University of São Paulo, Brazil. His area of research is abrasive machining process and its automation. He has published more than 200 papers and four patents. Currently he is the President of the Institute for Technological Research of the State of São Paulo and a member of the International Academy for Production Engineering.

Antje G. Orths

Energinet.dk
Electricity System Development
Fredericia, Denmark
ano@energinet.dk

Chapter C.17, Sect. 17.6

Antje Orths graduated in Electrical Engineering at the TU Berlin and received the PhD degree from the Otto-von-Guericke University Magdeburg, Germany. Afterwards she lead the Critical Infrastructures Group at the Fraunhofer-Institute IFF in Magdeburg. Since 2005 she has been with the Planning Department of Energinet.dk, the Danish TSO for electricity and natural gas. Her research interests are planning issues such as the implementation of wind energy into electric power networks and systems. She is a member of the IEEE, VDE-ETG, and CRIS.

Vince Piacenti

Robert Bosch LLC
System Engineering, Diesel Fuel Systems
Farmington Hills, MI, USA
vince.piacenti@us.bosch.com

Chapter B.10, Sect. 10.4

Vince Piacenti received the B.S. degree in Mechanical Engineering from Bradley University in 1976. He spent 7 years at Bosch's Headquarters working on diesel fuel injection. Currently he is Senior Manager at Bosch in Michigan responsible for hydraulic systems integration. He has experience in all forms of diesel and gasoline fuel injection, and research experience with thermoplastics and alternate fuels.

Jörg Pieschel

Otto-von-Guericke University
Institute of Materials and Joining Technology
Magdeburg, Germany
pieschel@mb.uni-magdeburg.de

Chapter B.7, Sect. 7.4.2

Jörg Pieschel earned the Dipl.-Ing. and and Dr.-Ing. degrees from Otto-von-Guericke University Magdeburg in 1991 and 1999, respectively. In 2001 he was promoted to Head of Laboratories of the Materials and Joining Technology Institute and Welding Supervisor of the University of Magdeburg. Dr. Pieschel is an International Welding Engineer (IWE) and has specialized on joinability of materials, innovation of laser joining processes, quality assurance, and clarification of welding damages.

Stefan Pischinger

RWTH Aachen University
Institute for Combustion Engines
Aachen, Germany
pischinger@vka.rwth-aachen.de

Chapter B.13, Sect. 13.1

Professor Pischinger studied Mechanical Engineering at RWTH Aachen University. Form 1985 to 1989 he worked as a Research Assistant at the Sloan Automotive Laboratory at MIT until he received the PhD degree for his work on spark ignition. Form 1989 to 1997 he held various positions at Daimler, working on diesel and gasoline engine development, and became project leader for the new Common Rail V8-Diesel engine. Since 1997 he has been Professor at RWTH Aachen University and Director of the Institute for Combustion Engines. Since April 2003 he has also been President and CEO of FEV, an Engineering Services Company in the field of combustion engines and powertrain.

Didier M. Priem

École Centrale Nantes
Department of Materials
Nantes, France
didier.priem@ec-nantes.fr

Chapter B.7, Sect. 7.4.9

Didier M. Priem is an Engineer and Laboratory Manager for welding joining and forming technologies. For the last 20 years, he has been actively engaged in electromagnetic and electrohydraulic forming and welding technologies. He holds a patent on the electromagnetic forming of titanium dental parts.

Frank Riedel

Fraunhofer-Institute for Machine Tools and Forming Technology (IWU)
Department of Joining Technology
Chemnitz, Germany
frank.riedel@iwu.fraunhofer.de

Chapter B.7, Sect. 7.4.5

Frank Riedel graduated as a Mechanical Engineer, receiving the PhD degree and PhD habilitatus in Joining Technology from Chemnitz University of Technology. From 2003 to 2006 he was the Deputy Head of Professorship of Welding Technology and Chief Engineer of the Institute of Manufacturing/Welding Technology at Chemnitz University of Technology. Since 2006 he has been establishing the Department of Thermal Processing at the Fraunhofer-Institute for Machine Tools and Forming Technology in Chemnitz, Germany.

Holger Saage

University of Applied Sciences of Landshut
Faculty of Mechanical Engineering
Landshut, Germany
holger.saage@fh-landshut.de

Chapter B.3, Sect. 3.7

Holger Saage received the PhD degree from the Otto-von-Guericke University Magdeburg. His experience covers various fields of materials science, including alloy development, powder technology, modeling of mechanical properties, and methods of materials analysis. Prior to joining the university he worked as a Materials Analyst in the semiconductor industry (AMD Saxony, Dresden). His current research work focuses on high-temperature materials, especially those with intermetallic phases such as TiAl and Mo silicides.

Shuichi Sakamoto

Niigata University
Department of Mechanical and Production Engineering
Niigata, Japan
sakamoto@eng.niigata-u.ac.jp

Chapter B.8, Sect. 8.2

Shuichi Sakamoto worked as Research Associate of the JSPS in 1989 and 1990. After that he received the PhD degree from Niigata University in 1991. He joined Niigata University in 1991 as a Research Associate, and became Associate Professor in 1998. His research interests are the development of new measuring or detection methods using acoustics, the characteristics of airborne sound-absorbing materials, and noise control.

Roger Schaufele

California State University
Long Beach, CA, USA
rdschaufele@aol.com

Chapter B.13, Sect. 13.4

Roger Schaufele graduated from Rensselaer Polytechnic Institute in 1949 with a degree in Aeronautical Engineering and obtained a M.S. degree in Aeronautics from CalTech in 1952. During the following years he held a number of key roles including Lead Aerodynamics Engineer for the DC-8 project, Aerodynamicist on the DC-9 and DC-10, and General Manager of Commercial Advanced Products. Since retiring from Douglas in 1989, he has been a part-time faculty member at California State University, Long Beach. He teaches courses in aircraft preliminary design and performance.

Markus Schleser

RWTH Aachen University
Welding and Joining Institute
Aachen, Germany
schleser@isf.rwth-aachen.de

Chapter B.7, Sect. 7.4.7

Dipl.-Ing. Markus Schleser studied at RWTH Aachen University and majored in Production Engineering in 2001. From 2001 to 2007 he has been employed as a Research Assistant in the Department of Adhesive Bonding and since August 2007 has been Chief Engineer at the Welding and Joining Institute (ISF) of RWTH Aachen University.

Meinhard T. Schobeiri

Texas A&M University
Department of Mechanical Engineering
College Station, TX, USA
tschobeiri@tamu.edu

Chapter B.12

Dr. Schobeiri, a Professor of Mechanical Engineering, received his entire engineering education at the Technical University Darmstadt, Germany. He was Group Leader for Gas Turbine Aero-Thermodynamic Design at Brown Boveri Co., Baden, Switzerland. His area of expertise includes unsteady aerodynamics, turbine and compressor aerodynamics design, and nonlinear gas turbine engine dynamic simulations. He is the author of one book and more than 100 technical papers and reports. He is a member of VDI and a Fellow of the ASME.

Miroslaw J. Skibniewski

University of Maryland
Department of Civil and Environmental Engineering
College Park, MD, USA
mirek@umd.edu

Chapter B.14, Sects. 14.1, 14.3–14.8

Prof. Skibniewski is the holder of the A. James Clark Endowed Chair in Construction Engineering and Project Management at the University of Maryland, USA. A recipient of Civil Engineering degrees from Warsaw University of Technology and from Carnegie Mellon University and author of over 150 publications, he serves as Editor-in-Chief of *Automation in Construction*, an international research journal.

Jagjit Singh Srai

University of Cambridge
Centre for International Manufacturing
Institute for Manufacturing
Cambridge, UK
jss46@cam.ac.uk

Chapter B.15, Sect. 15.3

Jag Srai is currently Head of the Centre for International Manufacturing, Institute for Manufacturing, University of Cambridge, an academic and practice-focused group of some 25 personnel. Before joining Cambridge University, Jag's previous Director-level roles have been in industry, working for leading multinationals in manufacturing and supply chain management, with over 17 years industrial experience. Jag has a degree in Chemical Process Engineering and is a Chartered Engineer and a Fellow of the Institute for Chemical Engineers.

Vivek Srivastava

Corporate Technology Strategy Services
Aditya Birla Management Corporation
Navi Mumbai, India
vivek.srivastava@adityabirla.com

Chapter B.3, Sect. 3.3

Vivek's research interest include room- and elevated-temperature mechanical properties of materials and their relationship with microstructure. He completed the PhD degree at the University of Sheffield, UK, on high-temperature creep of metals and alloys and was awarded the Bournton Medal for his thesis. He has published over ten research papers in international journals and has a deep interest in commercialization of technology.

Peter Stephan

Technical University Darmstadt
Institute of Technical Thermodynamics
Department of Mechanical Engineering
Darmstadt, Germany
pstephan@ttd.tu-darmstadt.de

Chapter B.4

Peter Stephan is a Professor at the Technische Universitaet Darmstadt and Head of the Institute of Technical Thermodynamics since 1997. His main fields of research are boiling heat transfer, microscale heat and mass transfer, interfacial phenomena, heat pipe technology, and drying and freezing processes. He is president of the VDI Heat and Mass Transfer Committee and is a member of several international thermal science associations.

Zbigniew A. Styczynski

Otto-von-Guericke University
Electric Power Networks and Renewable Energy Sources
Magdeburg, Germany
sty@ovgu.de or sty@ieee.org

Chapter C.17, Sects. 17.2, 17.7

Zbigniew Styczynski received the PhD degree from the Technical University of Wrocław, Poland. He worked at the University of Stuttgart, Germany, and in 1999 became Professor and Chair of Electric Power Networks and Renewable Energy Sources at the Otto-von-Guericke University, Magdeburg. From 2002 until 2006 he was the dean of the EE Faculty and since 2006 has been the President of the Centre of Renewable Energy Saxonia-Anhalt, Germany.

P.M.V. Subbarao

Indian Institute of Technology
Mechanical Engineering Department
New Delhi, India
pmvs@mech.iitd.ac.in

Chapter C.16

Dr. P.M.V. Subbarao is an Associate Professor of Mechanical Engineering Department at IIT, Delhi. He obtained the PhD degree in Mechanical Engineering from IIT Kanpur in 1996. He has developed ultramicro hydropower plants installed in remote rural areas for rural electrification. He has also developed a technology for production of bio-CNG from biogas for use in automobiles. His current research activities include power generation, renewable energy systems, and micro and pico power-generation systems.

Oliver Tegel

Dr.-Ing. h.c. F. Porsche AG
R&D, IS-Management
Weissach, Germany
oliver.tegel@porsche.de

Chapter B.13, Sect. 13.2

Dr. Oliver Tegel is working at IS-Management in R&D at Porsche AG. Previously, he has managed business process reengineering projects related to various aspects of the car development process. His educational background is in engineering design and design methodology, in which he earned the Doctor degree from the Technical University Berlin, Germany.

A. Erman Tekkaya

ATILIM University
Department of Manufacturing Engineering
Ankara, Turkey
etekkaya@atilim.edu.tr

Chapter B.7, Sect. 7.2

Professor Tekkaya completed his doctoral thesis at the University of Stuttgart in 1985. Until 2005 he was Professor at the METU, Ankara. From 2005 he was the Chairman of the Department of Manufacturing at ATILIM University in Ankara, Turkey. From 2007 he was Director of the Institute of Forming Technology and Lightweight Construction. His current research is on numerical simulation of sheet/bulk metal forming processes and material characterization. He is a member of CIRP and President of the International Cold Forging Group (ICFG).

Klaus-Dieter Thoben

University of Bremen
Bremen Institute for Production and Logistics GmbH Department of ICT Applications in Production
Bremen, Germany
tho@biba.uni-bremen.de

Chapter B.15, Sect. 15.10

After finishing his studies in Mechanical Engineering, Klaus-Dieter Thoben worked as a Research Assistant at the Department of Production Engineering at the University of Bremen, where he received the Doctor of Engineering degree in 1989. He received the state doctorate (Habilitation) in the domain of Production Systems in 2002. In the same year he was appointed Professor for IT Applications in Production Engineering at the University of Bremen. Since 2003 he has been Director of Bremen Institute for Production and Logistics GmbH (BIBA) in the Department of ICT Applications in Production.

Marcel Todtermuschke

Fraunhofer-Institute for Machine Tools and Forming Technology
Department of Assembling Techniques
Chemnitz, Germany
marcel.todtermuschke@saxonia.net

Chapter B.7, Sect. 7.4.5

After studying mechanical engineering with major subject welding and manufacturing technologies, Marcel Todtermuschke started working at the University of Technology Chemnitz in the Department of Welding Technology. He worked in various fields such as welding design and finally gained his Doctorate in the field of mechanical joining. Since 2006 he has been concentrating on assembling technologies at the Fraunhofer-Institute for Machine Tools and Forming Technology.

Helmut Tschoeke

Otto-von-Guericke University
Institute of Mobile Systems
Magdeburg, Germany
tschoeke@ovgu.de

Chapter B.10

Helmut Tschoeke is a Professor of Reciprocating Machines at the University of Magdeburg. From 1981 to 1995 he worked with Bosch Diesel Division, where he was responsible for development and production of distributor and inline pumps. During his career he has held positions as Chief Engineer and Executive Plant Manager. He is an active member of VDI and member of SAE and the head of the automotive research program at the University of Magdeburg.

Jon H. Van Gerpen

University of Idaho
Department of Biological and Agricultural Engineering
Moscow, ID, USA
jonvg@uidaho.edu

Chapter B.10

Jon Van Gerpen has been Professor and Department Head of Biological and Agricultural Engineering at the University of Idaho since July 2004. Before that, he was a Professor of Mechanical Engineering at Iowa State University for 20 years. He received the PhD degree from the University of Wisconsin-Madison in 1984. His current research interests include the production and utilization of biofuels and the development of a nationwide biodiesel education program.

Anatole Vereschaka

Moscow State University of Technology "STANKIN"
Department of Mechanical Engineering Technology and Institute of Design and Technological Informatics Laboratory of Surface Nanosystems
Russian Academy of Science
Moscow, Russia
dr_averes@rambler.ru

Chapter B.7, Sects. 7.3.1, 7.3.2

Anatoly Vereschaka, who received the PhD and Science Doctor degrees from Moscow State University in 1965 and 1986, respectively, is a Professor of Engineering Technology, Material Cutting Technology, and Surface Engineering Technology in the Department of Mechanical Engineering Technology of Moscow State University of Technology (STANKIN), Russia. His research interests are in the physics of metal cutting processes, design theory and methodology of wear resistance, and functional coating for cutting tools. He is Director of the Research Laboratory of Surface Nanosystems of the Russian Academy of Sciences.

Detlef von Hofe

Krefeld, Germany
detlef.von.hofe@web.de

Chapter B.7, Sect. 7.4.1

Detlef von Hofe was responsible for stationary gas turbine fabrication at Siemens Power Generation until 1991 and has been the Chairman of CEN/TC 121 Welding since 2003.
He was the Chief Executive Director and a Member of the Executive Council of the DVS, the German Welding Society until January 2006. He was appointed a Honorary Professor at the Otto-von-Guericke University in Magdeburg, teaching quality assurance in welding technology.

Nikolaus Wagner

RWTH Aachen University
ISF Welding and Joining Institute
Aachen, Germany
wa@isf.rwth-aachen.de

Chapter B.7, Sect. 7.4.4

Dipl.-Ing. Nikolaus Wagner studied Mechanical Engineering from 1997 to 2004 at the RWTH Aachen University with an emphasis on production engineering. Since January 2005 he has been working as a Scientific Assistant in the Laser Beam Department of the ISF – Welding and Joining Institute at the RWTH Aachen University. After carrying out research on laser hybrid welding of light metals, he is currently working on laser-beam welding with pressure-sensitive adhesives. He is also doing research in the Cluster of Excellence Integrative Production Technology for High-Wage Countries.

Jacek G. Wankowicz

Institute of Power Engineering
Warsaw, Poland

Chapter C.17, Sect. 17.1

Jacek Wankowicz received the PhD and Dr. Sc. degrees from the Technical University of Wrocław, Poland. He worked at the Technical University of Wrocław, Poland and Bayero University Kano, Nigeria, and since 1997 has been the Managing Director of the Institute of Power Engineering, Warsaw, Poland. He is a Member of many Supervisory Boards, Member of the IEC, CIGRE and is currently the President of the Polish National Committee of the CIGRE.

Ulrich Wendt

Otto-von-Guericke University
Department of Materials and Joining Technology
Magdeburg, Germany
wendt@ovgu.de

Chapter B.3, Sects. 3.2, 3.7

Ulrich Wendt studied chemistry and received the Dr. rer. nat. degree for the investigation of a catalytic reaction and a second degree (Habil.) for work on polymer melt crystallization. He heads the Microscopy and Stereology Laboratory and lectures on microscopy, spectroscopy, image analysis, and chemical analysis. His research activities are related to microstructure characterization, topometry, and failure analysis.

Steffen Wengler

Otto-von-Guericke University
Faculty of Mechanical Engineering
Institute of Manufacturing Technology and Quality Management
Magdeburg, Germany
swengler@ovgu.de

Chapter B.8, Sect. 8.2

Steffen Wengler graduated and obtained the Doctor degree in Mechanical Engineering, both from Otto-von-Guericke University of Magdeburg, Germany. His fields of specialization include manufacturing measurement technology and gear metrology, mainly cylindrical involute gears, and gear pairs. Since 1990 he has been Head of the Laboratory for Measurement Technology in the Institute of Manufacturing Technology and Quality Management at this university.

Bernd Wilhelm

Volkswagen AG
Sitech Sitztechnik GmbH
Wolfsburg, Germany
bernd.wilhelm2@volkswagen.de

Chapter B.15, Sect. 15.7

Bernd Wilhelm has had a career in the automotive industry, first as Specialist for manufacturing strategies, later as Manager of a Volkswagen assembly line, as Chief Industrial Engineer for the Volkswagen Group, and as Executive Manager of plants in Zaragoza and Brussels. As CEO for Sitech, a subsidiary of VW, he guarantees global seat production for the Volkswagen group and teaches operations planning as a Professor. He chairs leading automobile associations and is a Board Member of the German MTM Society.

Patrick M. Williams

Assystem UK
Bristol, UK
pwilliams@assystemuk.com

Chapter B.15, Sect. 15.8

Patrick has been involved in the UK aerospace industry for over 25 years. He qualified with the MSc degree in Technology Management from Universities in Bristol. His career has covered managing programmes in the UK, Europe, and the USA. Recently, Patrick has been involved in European research work in collaborative engineering in aerospace, focusing on management techniques in the supply chain.

Lutz Wisweh

Otto-von-Guericke University
Faculty of Mechanical Engineering
Institute of Manufacturing Technology and Quality Management
Magdeburg, Germany
lutz.wisweh@ovgu.de

Chapter B.8, Sect. 8.1

Lutz Wisweh received the Dipl.-Ing. and Dr.-Ing. degrees from the University of Magdeburg. In 1999 he was a Visiting Professor at the Niigata University, Japan. In 1999 he had an unlimited Visiting Professorship at the Universidad Central de Las Villas, Cuba. Currently, he is an Extracurricular Professor at the University of Magdeburg. His research interests are mainly in the use of statistical methods in quality management and measurement uncertainty in manufacturing measurement technology.

Johannes Wodara

Schweißtechnik-Consult
Magdeburg, Germany
johanneswodara@hotmail.com

Chapter B.7, Sect. 7.4.6

Professor Wodara was working for many years in speciality areas of welding. He was Head of the Institute of Assembling and Welding Engineering at the Otto-von-Guericke University Magdeburg. He also taught fundamentals of manufacturing in vehicle and aircraft engineering at Hamburg University of Applied Sciences. He has more than 170 publications to his name and is (co)author of several books in the field of welding, soldering, and brazing.

Klaus Woeste

RWTH Aachen University
ISF Welding and Joining Institute
Aachen, Germany
wo@isf.rwth-aachen.de

Chapter B.7, Sect. 7.4.3

Dr.-Ing. Klaus Woeste studied Mechanical Engineering from 1994 to 2001 at RWTH Aachen University with emphasis on Production Engineering. From 2001 to 2004 he worked as a scientific assistant at the ISF – Welding and Joining Insitute at the RWTH Aachen University in the Electron Beam Department. From 2004 to August 2007 he was Chief Engineer of the ISF and was responsible for the coordination of scientific projects. He earned his doctorate in 2005. Since August 2007 he has been working in the Innovation Centre R&D of Gottwald Port Technology GmbH, Düsseldorf, Germany

Hen-Geul Yeh

California State University
Department of Electrical Engineering
Long Beach, CA, USA
heyeh@csulb.edu

Chapter A.2

Dr. Hen-Geul Yeh's research areas are dynamics and adaptive controls, digital signal processing, and digital communication systems. He has been selected twice as the NASA summer faculty fellow, in 1992 and 2003. He was the recipient of four NASA tech brief awards and one NASA new technology award. He was the recipient of the Aerospace Corporation inventor's award. He has been selected as the Boeing A. D. Welliver faculty summer fellow in 2006. He owns several engineering patents.

Hsien-Yang Yeh

California State University Long Beach
Department of Mechanical and
Aerospace Engineering
Long Beach, CA, USA
hyyeh@csulb.edu

Chapter A.2

Hsien-Yang Yeh received the PhD degree from the University of Southern California. His research interests include the mechanics of composite materials, fracture mechanics, structural and machine components failure analysis, and nanotechnology applications. Dr. Hsien-Yang Yeh's research interests include the mechanics of composite materials, fracture mechanics, structural and machine components failure analysis, and nanotechnology applications.

Shouwen Yu

Tsinghua University
School of Aerospace
Beijing, P.R. China
yusw@mail.tsinghua.edu.cn

Chapter A.2

Professor Shouwen Yu is a Professor in the Department of Engineering Mechanics at Tsinghua University, Beijing, China. Professor Yu has been working in the field of fracture mechanics and nano/micro/mesomechanics over the last few decades. From 1985 to 1987 he was with the Institute of Mechanics, Technische Hochschule Darmstadt, Germany as a Visiting Research Fellow under an Alexander von Humboldt Fellowship. Professor Yu was Vice President of Tsinghua University from 1992 to 1999 and Dean of Graduate School of Tsinghua University from 1994 to 1999.

Detailed Contents

Part C Complementary Material for Mechanical Engineers

Subject Index

C

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

X

Y

Z